# Comparing different crowdsourced data: an analysis of Flickr elements, qualities and activities with Geo-Wiki land cover

Comber, Alexis ; Purves, Ross S

40

# Extended Abstract Proceedings of the

# GIScience 2014

**Compiled and Edited by**

**Kathleen Stewart**
**Edzer Pebesma**
**Gerhard Navratil**
**Paolo Fogliaroni**
**Matt Duckham**

Department of Geodesy and Geoinformation
Vienna University of Technology


Geo
info

# Extended Abstract Proceedings of the GIScience 2014

**Compiled and Edited by by**
**Kathleen Stewart**
**Edzer Pebesma**
**Gerhard Navratil**
**Paolo Fogliaroni**
**Matt Duckham**

Department of Geodesy and Geoinformation
Vienna University of Technology
Gusshausstraße 27-29/120.2
1040 Vienna, Austria

Series Editor

Andrew U. Frank
Dept. of Geodesy and Geoinformation
Vienna University of Technology
Gusshausstr. 27-29/120.2
A-1040 Vienna, Austria
frank@geoinfo.tuwien.ac.at

ISBN 978-3-901716-42-3 GeoInfo Series Vienna


Printed in Austria

Typesetting: Camera ready by author/editor
Printing and binding: Hochschülerschaft, TU Vienna
Wirtschaftsbetriebe GmbH

# Preface

In 2014, the 8th International Conference on Geographic Information Science (GIScience, www.giscience.org) was hosted by the Vienna University of Technology. This volume contains the extended abstracts accepted for oral and poster presentation. The GIScience conference series was created as an exchange platform for researchers interested in advancing fundamental aspects of Geographic Information Science. The first conference was held in Savannah, Georgia, USA, in 2000 and since then it was organized bi-annually. After three successful conferences in the USA, the first conference outside the USA was held in Münster, Germany. Since then, GIScience was organized alternately in the USA and in Europe and after Münster, Germany (2006) and Zürich, Switzerland (2010) the meeting in Vienna is the third meeting in Europe.

84 full papers (up to 15 pages) and 155 extended abstracts (up to 1500 words and 6 pages) were submitted to the conference. Full papers were supposed to describe results of scientific work whereas extended abstract to present work in progress. All full papers went through a thorough review by at least 3 members of the program committee and 23 of them were accepted for oral presentation and published in a Volume of Springer's Lecture Notes in Computer Science. The extended abstracts were screened by at least two reviewers from the program committee. 52 of them were accepted for short oral presentations and 61 were presented at a poster session. The extended abstracts are published in this book. Since several authors withdrew their extended abstracts after acceptance, only 101 extended abstracts are included in the book.

GIScience brings together experts from academia, industry, and government organizations. The range of topics covered by extended abstracts in this book is impressive: Classical topics like spatial analysis and land cover and land use classification, uncertainty, decision-support, or spatial relations are still represented well but complemented by hot topics, e.g., time and spatio-temporal structures or user-generated content, or big dada.

In the preparation of the conference the GIScience community received bad news: Prof. Peter Fisher died on 20 May 2014. He was (co-)author of 2 full papers and 4 extended abstracts submitted to GIScience 2014 and was a valuable member of the program committee since the start of the conference series. Two weeks later, on 3 June 2014, the program committee lost Prof. Carolyn Merry. She was former president of the US-based UCGIS and past President of ASPRS. Both colleagues will be missed.

As program co-chairs, workshop and tutorial chair, and local organizer we want to thank the many people who made this conference possible: The authors who submitted their work, the program committee and the additional reviewers who guaranteed the quality of the scientific content, the co-organizer Austrian Computer Society and specifically Christine Haas, the staff of the Research Group

Geoinformation at the Department for Geodesy and Geoinformation, and the Vienna University of Technology. Special thanks go to Eva-Maria Holy for her invaluable and constant support in the whole process of conference organization.

August 2014

Kathleen Stewart
Edzer Pebesma
Gerhard Navratil
Paolo Fogliaroni
Matt Duckham

# Organization Committee

## General Chair

Andrew U. Frank, Vienna University of Technology

## Program Co-Chairs

Matt Duckham, University of Melbourne
Kathleen Stewart, University of Iowa
Edzer Pebesma, University of Münster

## Workshop & Tutorial Chair

Paolo Fogliaroni, Vienna University of Technology

## Local Organizers

Gerhard Navratil, Vienna University of Technology
Eva-Maria Holy, Vienna University of Technology

## Program Committee

Ola Ahlqvist, Ohio State University
Luc Anselin, Arizona State University
Marc Armstrong, University of Iowa
Kate Beard-Tisdale, University of Maine
Scott Bell, University of Saskatchewan
Itzhak Benenson, Tel Aviv University
David Bennett, University of Iowa
Michela Bertolotto, University College Dublin
Ling Bian, University at Buffalo
Thomas Bittner, SUNY at Buffalo
Dan Brown, University of Michigan
Dirk Burghardt, Technical University of Dresden
Barbara Buttenfield, University of Colorado Boulder
Gilberto Camara, INPE, Brazil
Adrijana Car, University of Salzburg & German University of Technology in Oman
Nicholas Chrisman, RMIT University
Christophe Claramunt, Naval Academy Research Institute (NARI)
Keith Clarke, University of California, Santa Barbara
Eliseo Clementini, University of L'Aquila
Tom Cova, University of Utah
Isabel Cruz, University of Illinois at Chicago

Amy Lobben, University of Oregon
Alan MacEachren, Pennsylvania State University
William Mackaness, University of Edinburgh
Jeremy Mennis, Temple University
Carolyn Merry, Ohio State University
Harvey Miller, The Ohio State University
Alan Murray, Arizona State University
Tomoki Nakaya, Ritsumeikan University
Atsuyuki Okabe, University of Tokyo
Antonio Paez, McMaster University
Dimitris Papadias, UST, Hong Kong
Karin Pfeffer, University of Amsterdam
Dieter Pfoser, George Mason University
Alenka Poplin, HafenCity Universität Hamburg
Lilian Pun, Hong Kong Polytechnic University
Ross Purves, University of Zürich
Martin Raubal, ETH Zürich
Tumasch Reichenbacher, University of Zürich
Femke Reitsma, University of Canterbury
Maria Andrea Rodriguez-Tastets, Universidad de Concepción
Anne Ruas, Institut Franais des Sciences et Technologies des Transports
Christoph Schlieder, University of Bamberg
Falko Schmid, University of Bremen
Nadine Schuurman, Simon Fraser University
Shih-Lung Shaw, University of Tennessee
Takeshi Shirabe, Royal Institute of Technology Sweden
Bettina Speckmann, Eindhoven University of Technology
Emmanuel Stefanakis, University of New Brunswick
Paul Sutton, University of Denver
Jean-Claude Thill, University of North Carolina at Charlotte
Sabine Timpf, University of Augsburg
Paul Torrens, University of Maryland
Ming-Hsiang Tsou, San Diego State University
Marc van Kreveld, Utrecht University
Monica Wachowicz, University of New Brunswick
Jan Oliver Wallgrün, Pennsylvania State University
Shaowen Wang, University of Illinois
Robert Weibel, University of Zürich
John Wilson, University of Southern California
Stephan Winter, The University of Melbourne
Michael Worboys, University of Greenwich
Dawn Wright, ESRI

Ningchuan Xiao, Ohio State University
Yichun Xie, Eastern Michigan University
Chaowei Yang, George Mason University
Bailang Yu, East China Normal University
May Yuan, University of Oklahoma

## Additional Reviewers

Roger Bivand, Norway
Olivier Bonin, France
Min Chen, Hong Kong
Matthew Dube, USA
Junchuan Fan, USA
Riccardo Fellegara, Italy
Eric Fung, China
Jon Goergen, USA
Venkat Raghavan Ganesh, USA
Peng Gao, USA
Stefan Hahmann, Germany
Pierre Hallot, USA
Paul Hiemstra, Netherlands
Mingyuan Hu, Hong Kong
David Jonietz, Germany
George Kellaris, China
Patrick Laube, Switzerland
Joshua Lewis, USA
Xuecao Li, China
Silvia Nittel, USA
Mark Padgham, Germany
Christoph Stasch, Germany
Lu Tan, China
Maria Vasardani, Australia
Emily White, USA

# TABLE OF CONTENTS

## ABSTRACTS FOR PRESENTATION

### Decision Support



### Geodemographic Classification



### Land Cover and Land Use

## Movement and Context



## Movement and Flows



## Space and Time



## Spatial Algorithms and Models

## Spatial Analysis: Activity



## Spatial Analysis: Inference



## Spatial Analysis: Patterns



## Spatial Perceptions

## Spatial Regions and Relations



## Spatiotemporal Structure



## Text

## Time Geography



## Uncertainty



## User-Generated Content

## ABSTRACTS FOR POSTERS

# Embodied decision making with animations

S. Maggi and S. I. Fabrikant

University of Zurich, Department of Geography, Winterthurerstr. 190, CH –8057 Zurich, Switzerland.
Email: {sara.maggi, sara.fabrikant}@geo.uzh.ch

## 1. Introduction

Map animations have become popular devices to depict complex spatio-temporal phenomena. Intuitively, animations seem to be an ideal choice to depict moving objects over time (e.g., aircraft movements over an airport at a particular day), because real-world object movements are congruently depicted with graphical objects moving in a display (Tversky 2002). Visuo-spatial decision-making might not only be influenced by external stimuli (e.g., the perceptual salience and thematic relevance of visual cues), but also by a viewer's internal emotional state (e.g., mood or motivation) (Koelstra 2012). To gain more insights on users' perceptual, cognitive and emotional processes, we propose a long-term empirical research framework to empirically assess dynamic visuo-spatial displays, based on triangulation that couples eye movement data with electrodermal activity (EDA), electroencephalography (EEG), and traditional questionnaires (Holmqvist et al. 2011; Maggi and Fabrikant 2014).

We present a subset of preliminary results of a human-subject experiment in the context of air traffic control to monitor aircraft movements. We aim to investigate how display- (i.e., animation types), data- (i.e., characteristics of the depicted objects), and user-related factors (i.e., individual/group differences) might influence visuo-spatial decision-making with animations.

## 2. Methods

Standard air traffic control (ATC) displays typically show aircraft movements with semi-static animations in which aircraft positions are refreshed every 4s. Following a mixed factorial design, we set out to investigate how the independent variables display design (i.e., semi-static vs. continuous aircraft movements), aircraft movement dynamics and context (i.e., varying speeds and number of displayed aircrafts), and user characteristics (i.e., ATC expertise, spatial ability, and psycho-physiological state) might influence the accuracy and speed of aircraft movement detection (i.e., dependent variables). Figure 1 shows a static version of a stimulus, inspired by French air traffic control (ATC) radar screens, representing four aircrafts moving at different speeds in the same direction.

In a between subject design, we asked eighteen air traffic controllers at the Ecole Nationale de l'Aviation Civile (ENAC) in Toulouse (e.g., ATC experts), and nineteen psychology students at Temple University in Philadelphia (e.g., ATC novices) to watch sixteen semi-static and sixteen continuous animations (N=32), and to click on the accelerating aircraft as soon as they detected it. Animations were presented digitally on a color monitor at 1920x1200 spatial (pixel) resolution. Participants had the possibility to stop the animation, once they identified the accelerating aircraft. On average, the animated portion of the experiment took about 16 minutes. Before and after being shown the animated displays, participants filled in a Short Stress State Questionnaire (SSSQ) that captures three individual factors of subjective stress, i.e., task engagement, distress, and worry (Helton et al. 2004). We recorded participants' eye movements, electrodermal activity, and EEG during the entire experiment.

Figure 1. Sample stimulus with four aircrafts moving from left to right.

## 3. Results and Discussion

We present a subset of our results, focusing on the dependent variables: response accuracy (Maggi and Fabrikant 2014) and EDA, across the independent variables animation design (i.e., semi-static vs. continuous animations), and users' ATC expertise levels (i.e., experts vs. novices). Figure 2 graphs response accuracy across animation type and expertise.

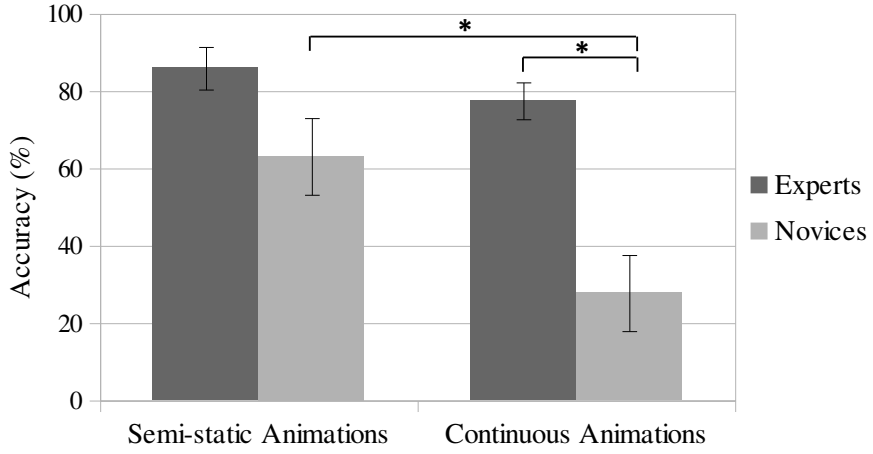



Figure 2. Mean response accuracy for ATC experts and novices across animation design conditions (error bars show standard error).

On average, response accuracy of novices is significantly higher ($F(1,17)=6.38$, $p<.022$) with semi-static displays (63%), compared to continuous displays (28%). Irrespective of display design, experts are more accurate than novices (i.e., close to 80%), but this difference is only significant for the continuous animation condition ($F(1,17)=22.19$, $p<.000$). We further analyzed participants' arousal response intensity that might have affected response accuracy. As shown in Figure 3 below, we compare the standardized, average area (i.e., integral) bounded by the SCR curve of the

phasic EDA by trials, across expertise level and animation design conditions (Boucsein 1992; Lykken and Venables 1971).

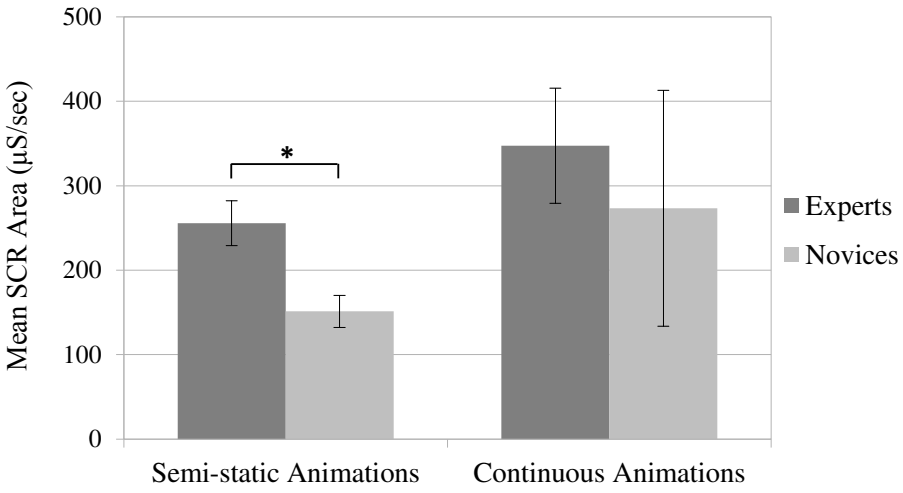
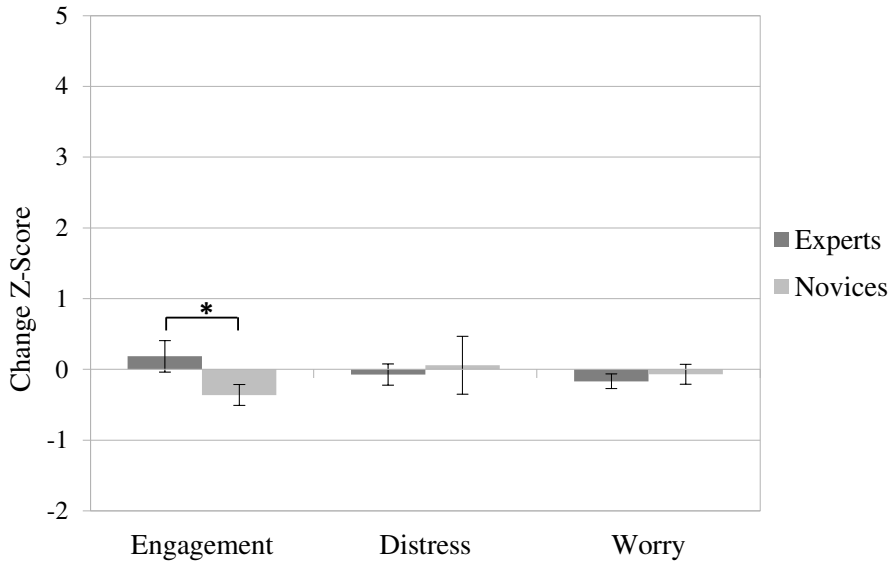



Figure 3. Mean area bounded by the SCR curve for ATC experts and novices across animation design conditions.

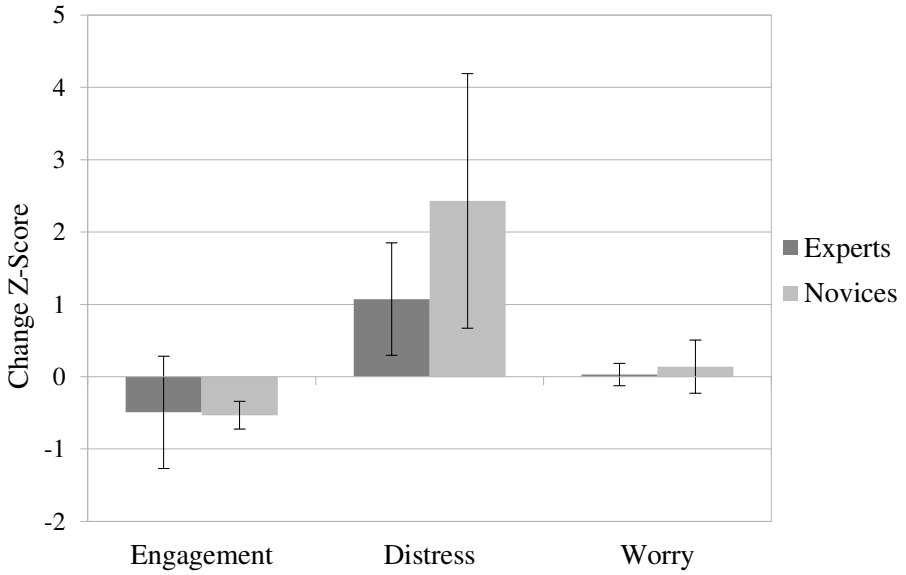
We only analyzed a subset of 20 participants to date (13 experts and 7 novices), due to noise in the recorded SCL data. We find that on average experts show higher arousal levels than novices (Figure 3). However, this arousal response density difference is only significant for the semi-static animations, due to large variances in the continuous animation condition. We additionally investigated participants' stress factor states that might further explain SCR differences, i.e., engagement, distress and worry, captured with the SSSQ (Helton et al. 2004). Figure 4 details change z-scores across animation type and expertise level. Experts show more engagement and less distress and worry with the familiar semi-static animations, compared to novices (Figure 4a). The difference is only barely significant for the factor engagement ($t(16)=3.00$, $p<.050$), probably again because of the large variance in the novice population. This picture is almost mirrored for the continuous animation condition (Figure 4b). Now experts and novices show similar low engagement patterns, but novices exhibit more distress and worry than experts. None of the SSSQ score differences are significant in the continuous animation condition (Figure 4b), however. This might again be due to the large variance in the novice group. Significant differences both found in arousal levels and SSSQ scores between experts and novices for the semi-static animation condition might have a relevant relationship that needs further investigation.

While experts perform the experimental task more accurately than novices overall, irrespective of the animation design, they appear to have been less engaged, and more distressed with the unfamiliar continuous animation type. Higher distress levels coupled with worry might explain novices' even lower response accuracy with the continuous animations, compared to those of the experts. Conversely, being less distressed and less worried with the semi-static condition, had a positive effect on response accuracy for novices, compared to their low performance in the continuous animation condition. In fact, anecdotal evidence from open-ended questionnaires suggests that novice participants found it much more difficult to solve tasks with the continuous

animations than with the semi-static ones. Familiarity with a display (i.e., training) seems to have a greater effect on decision-making accuracy than novelty of display design.




(a) Semi-static animations.




(b) Continuous animations.

Figure 4. SSSQ change z-scores across expertise across animation types.

## 4. Summary and Outlook

We presented preliminary results of a first study within a novel empirical, embodied research framework to assess animated visual displays based on psycho-physiological data triangulation. Our results confirm previous animation studies suggesting that prior knowledge and display design influence visuo-spatial decision-making with animated graphical displays (Kriz and Hegarty 2007). By specifically quantifying participants' arousal and stress factor states, analyzed across expertise levels, we are able to more deeply investigate how cognitive, perceptual, and psycho-physiological processes might affect the effectiveness of animation design in the context of air traffic control. We intend to further include collected EEG data streams and eye movement recordings in our analysis to better disentangle emotional, cognitive and perceptual factors from display design and task contexts. We will also further investigate event-related electrodermal activity, as there is a known positive relationship between physiological arousal and task performance (Yerkes and Dodson 1908).

We are currently designing follow-up experiments to further investigate how the inclusion of contextual information (e.g., a wind map), and the increase of perceptual salience (e.g., through visual variables) might affect decision-making with animated displays.

With the empirical results of this user-centered empirical research agenda we hope to develop general design guidelines for perceptually salient, affectively engaging, and cognitively inspired animations that support effective and efficient visuo-spatial exploration and decision-making of spatio-temporal phenomena and processes.

## References

Boucsein W, 1992, Electrodermal Activity. Berlin, Germany: Springer.

Helton WS, 2004, Validation of a Short Stress State Questionnaire. Proceedings of the 48th Meeting of the Human Factors and Ergonomics Society, 48(11): 1238-1242.

Holmqvist K et al., 2011, Eye Tracking: A comprehensive Guide to Methods and Measures. Oxford: Oxford University Press.

Koelstra S et al., 2012, Deap: A Database For Emotion Analysis Using Physiological Signals. IEEE Transactions on Affective Computing 3(1): 18-31.

Kriz S and Hegarty M, 2007, Top-Down and Bottom-Up Influences on Learning from Animations. International Journal of Human-Computer Studies, 65(11): 911-930.

Lykken DT and Venables PH, 1971, Direct Measurement of Skin Conductance: A Proposal for Standardization. Psychophysiology,8(5): 656-672.

Maggi S and Fabrikant SI, 2014, Assessing Animated Air Traffic Control Displays. Abstract proceedings of CARTOCON 2014, Olomouc, Czech Republic, Feb. 25-28, 2014: 20.

Tversky B et al., 2002, Animation: Can It Facilitate? International Journal of Human Computer Studies, 57(4): 247-262.

Yerkes RM and Dodson JD, 1908, The Relation of Strength of Stimulus to Rapidity of Habit-Formation. Journal of Comparative Neurology and Psychology, 18: 459–482.

## Acknowledgments

We thank participants from ENAC (France) and Temple University (USA) who participated in our experiment. We are also grateful to Christophe Hurter, Jean-Paul Imbert and Maxime Cordeil at ENAC, including Tim Shipley and Kelly Bower at Temple for their invaluable collaboration, including theoretical and technical support in the design and realization of the study.

# Comparing VGI contributions across political units using geovisual analytics

S. Quinn[1]

[1]Department of Geography, The Pennsylvania State University, 302 Walker Building, University Park, PA 16802
Email: sdq107@psu.edu

## 1. Introduction and context

This paper describes a web-based geovisual analytics application for comparing volunteered geographic information (VGI) contributions between political units, such as countries. The research focuses on OpenStreetMap (OSM) for its worldwide utility and spatial scope; however, the approach could be applied to other sources of VGI.

VGI is growing in popularity and holds potential value for governments, emergency planners, and cartographers; however, it is still unclear whose world is represented with VGI and whether there are ways to calibrate and organize the "citizen sensors" that Web 2.0 has engendered (Goodchild 2007). For example, researchers have noted that although OSM is growing, "virtual black holes" persist due to personal biases, lack of technology infrastructure, cultural attitudes toward technology, and government policies restricting data collection or movement (Graham 2010, Zook and Graham 2007).

The tools described in this paper comprise the first phase of the author's longer-term theoretical and computational inquiry into the social textures imprinted in OSM across space and scale; in other words, when we look at a map comprised entirely of VGI, whose influence do we see? How many people are at work, how much of the influence can be classified as "local", and what social circumstances are affecting participation? An initial research objective of this effort is to develop geovisual analytics methods and tools that depict the geography of VGI.

Visually analyzing OSM in this way requires a foray into big data; namely, the "planet file" snapshots of the entire OSM database that are available for download in compressed XML format. Using the geometry of any political unit such as a country or city, the geographic features and contributor metadata can be clipped from the OSM planet file and summarized by computational and visual means.

## 2. Related work

Previous researchers have combed the OSM planet files as well as the full history "dump" of OSM data to summarize some aspects of contributor activity. For example, Neis and Zipf (2012) used the spatial coordinates of an individual user's edits to derive an "activity polygon" of the user's principal spatial domain in OSM. This polygon and other summary statistics can be obtained for any OSM user by consulting Neis' website (Neis 2014).

To visualize spatial aspects of VGI contributor activity, Roick et al. (2011) created an interactive map for exploring Europe's OSM data for patterns and anomalies. Variables such as date of latest edit, number of features, and number of attributes are depicted using hexagonal bins, and can be compared across temporal snapshots. The maps show stark contrast in contribution levels across borders and the urban-rural gradient, but offer no means for summarizing statistics for a particular country or identifying users most involved with mapping activity.

## 3. Methods

As the above literature demonstrates, the OSM planet files hold a wealth of information that can be mined using custom scripts. In this study, I combine OSM data with population figures to derive the following statistics that depict the health of OSM in a particular political unit:

1. Features/unit – The overall amount of mapping activity occurring
2. Features/capita – Mapping activity in the unit relative to the potential number of contributors in the unit
3. Features/area – How much content has been mapped relative to the potential features available for mapping
4. Contributors/unit – The raw amount of contributor interest in the unit, keeping in mind that a contributor may not necessarily live within the unit boundaries
5. Contributors/capita – The contributor interest relative to the number of people who would be expected to have some local knowledge of the unit
6. Contributors/area – The contributor interest relative to the area that must be covered
7. Features/contributor – How much the contributors tend to map
8. Contribution timestamps – How active the contributors have been over time

To visualize OSM activity across political units, I present a tool called Sud-OSM (based on the Spanish prefix for "South") that compares the above metrics for two countries of the user's choosing, side-by-side in an interactive web page. Sud-OSM is visible at: http://www.geovista.psu.edu/osm_in_south_america/index.html

Countries are chosen as a unit of study because data collection, data sharing, and human movement policies often vary across international boundaries. Furthermore, the OSM contributor wiki tends to organize work at the country level. South America is chosen as a study region because it exhibits a relatively low ratio of OSM members to population (Neis and Zipf 2012) and has not been systematically examined in the VGI literature.

Sud-OSM incorporates elements of geovisual analytics, wherein a user is invited to interact with the spatial and temporal dimensions of a large and heterogenous dataset in order to learn new things (Andrienko et al. 2007). Users of Sud-OSM select any two countries to compare and, in response, maps and computational summaries appear that are tailored to the user selections. Figure 1 shows the Sud-OSM user interface comparing the countries of Peru and Uruguay.

The side-by-side maps in Sud-OSM depict a sampling of OSM data in the selected countries. Users can move a slider to visualize the state of OSM at the end of any year from 2009 to 2013. Below the maps, users see tabs allowing the further exploration of the OSM snapshot from December 2013.

The "Overview" tab (visible in Figure 1) uses a series of histograms and bubble plots to summarize the eight OSM contribution variables listed above. The first histogram shows the raw number of editors and how many features they tended to add. The second histogram portrays the years the features were committed to OSM based on their timestamps.

The first bubble plot shows contributions and contributors normalized by country population, while the second bubble plot normalizes these figures by country area instead. Both bubble plots use total features in the unit to determine bubble diameter. A country appearing to the upper-right of the median has a greater proportion of contributors and contributions than we would expect for a country of its size or population. Users can optionally click a link to show all countries in one plot.

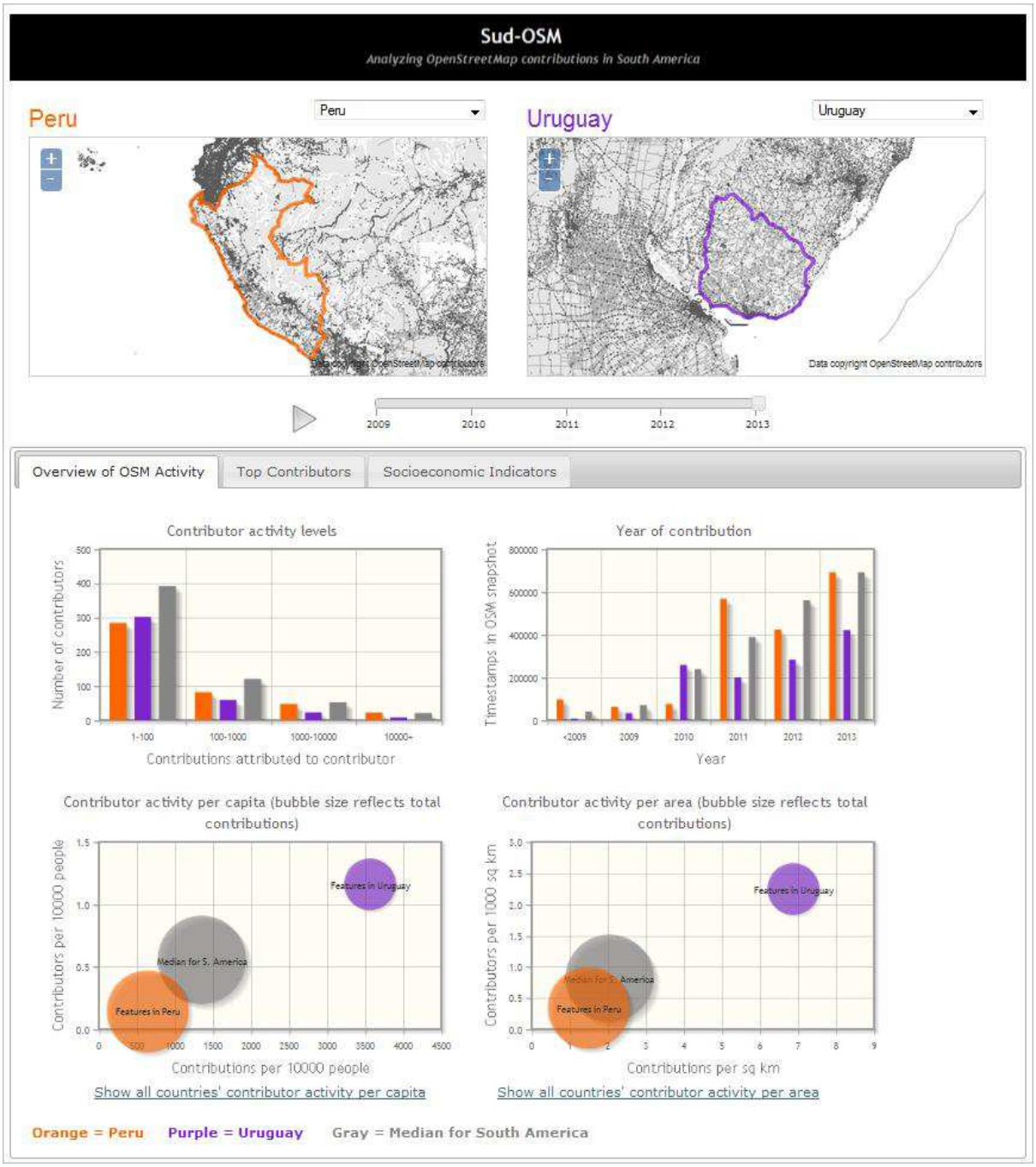



Figure 1. Sud-OSM user interface allowing comparison of two countries.

The "Top Contributors" tab (Figure 2) helps the analyst learn more about who is doing the most mapping. It lists the 20 most prolific contributors to OSM within the boundaries of each selected country. Clicking the name opens the personal OSM page for that user. Clicking the "details" link opens Pascal Neis' "How did you contribute" page summarizing the user's activity polygon and other contribution trends (Neis 2014). Taken together, these pages help reveal the influence of automated "bots", mass imports, and nonlocal mappers in OSM.

The "Socioeconomic Indicators" tab (Figure 3) offers contextual information about income and technology use in the compared countries. These charts are derived from IMF income figures and the Latin America Public Opinion Project (LAPOP) survey administered by Vanderbilt University. A comprehensive statistical analysis of these data and their

relations to OSM contributions is outside the scope of this paper; however, the charts still provide insight into the demographic context of OSM contributors in these countries.

## 4. Results

Exploration of the "Contributor activity levels" histogram in Sud-OSM confirms previous research by Neis and Zipf (2012) that most of the contributors to OSM only add a small number of features, whereas a small proportion contributes an enormous amount of edits. The "Top Contributors" list in Sud-OSM sheds light on the identities of these most active users. The lists for each country seem to be mostly unique, suggesting that editors with local interest are getting a foothold on the map; however, the lists contain various editors who appear to be from Europe, and at least one cleanup bot.


Figure 2. List of top contributors for selected countries.


Figure 3. Socioeconomic statistics for the selected countries.

The bubble plots indicate several countries consistently high in contributors and contributions relative to what would be expected for their size and populations. These include Uruguay and Chile. At the same time, some countries such as Brazil and Peru are consistently lower in these categories. The normalized figures might sometimes be difficult to guess by viewing the maps alone, especially with a country like Brazil that has a relatively high raw number of contributions.

The "Year of contribution" histogram reveals a year-over-year increase for the median number of features, but the individual countries are a different story. Some like Venezuela and Argentina have consistently risen, while others like Ecuador and Paraguay have experienced spikes and drops, perhaps indicating large data imports or the presence of a few very dedicated users who are no longer active.

## 5. Summary

Sud-OSM uses elements of geovisual analytics to compare OSM contributions across political units, helping researchers understand which countries have excelled or languished with accumulating VGI and therefore merit additional research. Although Sud-OSM focuses on countries in South America, the principles in Sud-OSM could be expanded to any geographic region or spatial unit of interest. For example, an analysis of metropolitan units might prove informative if political boundaries and population figures could be matched for each city.

## Acknowledgements

The author would like to thank Alan MacEachren for guidance on this project, Greg Milbourne for help with processing OSM datasets, and the Latin American Public Opinion Project (LAPOP) and its major supporters (the United States Agency for International Development, the United Nations Development Program, the Inter-American Development Bank, and Vanderbilt University) for making data available.

## References

Andrienko G, Andrienko N, Jankowski P, Keim D, Kraak MJ, MacEachren A, and Wrobel S, 2007, Geovisual analytics for spatial decision support: Setting the research agenda. International Journal of Geographical Information Science, 21(8), 839–857.

Goodchild MF, 2007, Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), 211–221.

Graham M, 2010, Neogeography and the palimpsests of place: Web 2.0 and the construction of a virtual earth. Tijdschrift Voor Economische En Sociale Geografie, 101(4), 422–436.

Latin American Public Opinion Project (LAPOP), 2014, The AmericasBarometer. Retrieved from http://www.lapopsurveys.org

Neis P, 2014, How did you contribute to OpenStreetMap? Retrieved from http://hdyc.neis-one.org/

Neis P, and Zipf A, 2012, Analyzing the contributor activity of a volunteered geographic information project—The case of OpenStreetMap. ISPRS International Journal of Geo-Information, 1(2), 146–165.

Roick O, Hagenauer J, and Zipf A, 2011, OSMatrix–grid-based analysis and visualization of OpenStreetMap. Proceedings of the 1st European State of the Map. Retrieved from http://koenigstuhl.geog.uni-heidelberg.de/publications/2011/Roick/Roick_2011_SotM.pdf

Zook MA, and Graham M, 2007, The creative reconstruction of the Internet: Google and the privatization of cyberspace and DigiPlace. Geoforum, 38(6), 1322–1343.

# Serious Games for Disaster Risk Reduction Spatial Thinking

Brian Tomaszewski[1], Jörg Szarzynski[2], David I. Schwartz[1]

[1]Rochester Institute of Technology, 152 Lomb Memorial Drive, Rochester, NY 14623, USA
Email: {bmtski; disvks}@rit.edu

[2]United Nations University Institute for Environment and Human Security, Platz der Vereinten Nationen 1,D-53113 Bonn, Germany
Email: szarzynski@ehs.unu.edu

## 1. Introduction

Spatial thinking is a learnable skill for structuring and solving spatial problems and decision making support, such as using a map to support navigation or structuring time using spatial metaphors (i.e. "the event is *far off* in the future"). The National Research Council (2006) defined *spatial thinking* as an amalgam of three items—concepts of space, tools of representation, and processes of reasoning (National Research Council 2006). *Serious games* target non-entertainment purposes using gaming concepts, such as a score based on actions taken for measuring game player learning and progress (Michael and Chen 2005).

In this paper, we present a theoretical framework and preliminary disaster risk reduction (DRR) serious spatial thinking game implementation. Our motivation for developing the framework and subsequent serious game is to:

1. Develop a new spatial thinking ability quantification method that can inform GIScience research, and
2. Implement scenarios and simulations that use commercial Geographic Information Systems (GIS) tools for novices to learn spatial thinking in a variety of application domains, such as DRR.

We are designing and developing our theoretical framework and current serious DRR spatial thinking game in close coordination and partnership with United Nations educational capacity development and DRR leaders. We believe that this connection demonstrates the societal relevance and broader impacts of our research (see United Nations University Institute for Environment and Human Security (n.d.)).

### 1.1 Theoretical Framework

Virtual geography-based games have seen growing research attention for teaching concepts, such as resource management and human-environment relations (Ahlqvist et al. 2012; Cheng et al. 2010). However, this prior work has not made an explicit focus using serious game concepts to measure and teach spatial thinking ability in novices (e.g., students) through industry-standard, commercial-grade GIS tools. A serious game would create a closer connection between the serious gaming experience and real-life (Ohmori et al. 2003). Figure 1 graphically summarizes our theoretical framework.

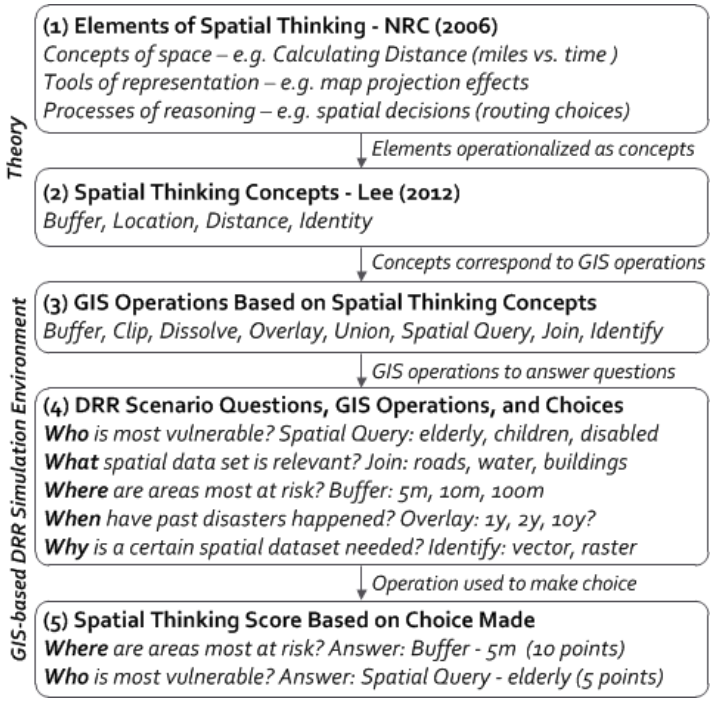



Figure 1: Serious Spatial Thinking Game Theoretical Framework. Parenthetical numbers in the text below refer to Figure 1 parenthetical numbers.

The framework starts with (1) the elements of spatial thinking defined in National Research Council (2006). Spatial thinking elements are then operationalized as (2) spatial thinking concepts based on Lee and Bednarz (2012) research to develop a spatial thinking ability test (STAT). For example, the spatial concept of a buffer can be considered a combination of the concepts of space and a tool of representation. Lee and Bednarz (2012) outlined tangible spatial thinking concepts specifically to measure spatial thinking ability via the STAT—concepts that we incorporate into our theoretical framework to justify quantifying spatial thinking ability. Many spatial thinking concepts, like buffer, correspond directly to (3) GIS operations found in industry-standard GIS tools such as ArcGIS[1]. Game players then match GIS operations, grounded in relevant spatial thinking concepts as defined in the literature, to develop (4) DRR serious game scenario questions and GIS operations to answer a given question. Finally, a score (5) based on spatial thinking choices made to allow the game player understand decisions made and the spatial thinking processes behind those decisions (Ohmori et al. 2003; Berse, Bendimerad, and Asami 2011).

## 2. Prototype Serious Spatial Thinking Game—Costal City Hurricane

Our prototype serious game was built inside ArcGIS using python GUI tools. Specific game scenario questions, GIS operations, and choices are encoded using JSON—see Blochel et al. (2013) for further game technical implementation details. Figure 2 shows the python-based game interface (bottom right) incorporated into ArcGIS for a coastal city hurricane scenario.

[1] www.arcgis.com

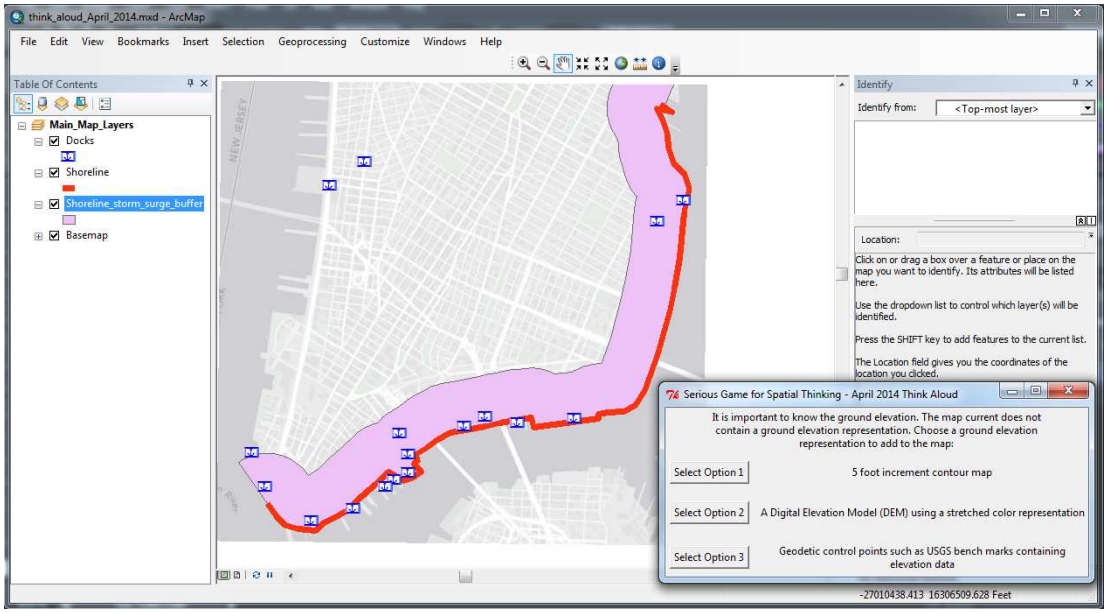
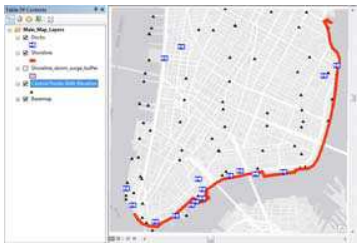



Figure 2: The serious spatial thinking game inside ArcGIS. In this example, the game player has just made a storm surge extent choice based on previous question choices (as seen by the purple buffer) and is now being prompted to select a ground elevation representation.

We are currently using a coastal city hurricane scenario in the game based on actual events from 2012 Hurricane Sandy and established hurricane disaster planning scenarios from the literature (US Department of Homeland Security 2006). In our game scenario, a Category 5 hurricane is approaching a coastal city and the game player must utilize spatial thinking supported by GIS to make choices (Figure 3a-c).

3a – Contour Map

3b – DEM



3c – Survey controls

Figures 3a-c: The three choices for the question shown in the bottom-right of Figure 2.

Figures 3a-c show three choices for determining the best elevation representation for thinking spatially about storm surge impact—a 5′ contour map (3a), a stretched color Digital Elevation Model (DEM) representation (3b), or geodetic control points containing elevation (3c). As per our theoretical framework, the line (contours), raster (DEM), and point (survey controls) aspects of the choices closely relate to spatial thinking component VIII on the STAT—"comprehending geographic features represented as point, line, or polygon" outlined in Lee and Bednarz (2012, 20). Furthermore, the choices are a good example of prompting novice spatial reasoning. An expert would know that a contour map would likely be the best choice of elevation representation to allow for ease of comparison with other layers via overlay. A DEM would be a second-best choice, and survey control points are the worst choice— as by themselves, they cannot easily convey elevation as a continuous surface.

### 2.2 Preliminary Results

We are currently using our serious spatial thinking game environment to gather evidence on novice spatial thinking supported with GIS tools via think-aloud sessions. Specifically, novices use the environment to answer a series of questions similar to those in Figures 2 and 3. They verbally express what they are thinking spatially about when using the environment, akin to prior spatial thinking/think-aloud research (Taylor and Tenbrink 2013). To date, we have tested seven college students (six graduate, one undergraduate) with an average two prior GIS classes (but no DRR experience) before participating in the think-aloud session, four of the seven are native English speakers and one is a deaf and hard of hearing. Select preliminary results indicate students with some GIS experience but unfamiliar with the DRR application domain are fairly capable at utilizing spatial thinking for different representation types such as points, lines, or polygons—"*with these choices* (about elevation), *you're really going to want to know how the water's going to flow in and where the water will go, so this last one* (the survey control points, Figure 3c) *doesn't have as much information as the first two*" (contour lines—Figure 3a and DEM—Figure 3b). However, more complex spatial thinking questions, such as choosing between clip, intersect, and union for determining priority evacuation areas revealed spatial thinking ability gaps — even after two GIS classes ("*I'm not that familiar with intersect*") and challenges with applying more advanced spatial thinking to the DRR domain — "*I think that where the storm surge buffer and the HAZMAT* (hazardous material) *buffer overlap is where you will want medical treatment…but I am blanking as to which* (GIS) *tool is the one best to use*..(for determining the medical treatment areas)."

## Acknowledgements

Portions of this work were supported through the National Science Foundation Science Master's Program "Decision Support Technologies for Environmental Forecasting and Emergency Response" (NSF DGE-1011458).

## References

Ahlqvist, Ola, Thomas Loffing, Jay Ramanathan, and Austin Kocher, 2012, Geospatial Human-environment Simulation through Integration of Massive Multiplayer Online Games and Geographic Information Systems. *Transactions in GIS,* 16 (3):331-350.

Berse, K.B., F. Bendimerad, and Y. Asami, 2011, Beyond geo-spatial technologies: promoting spatial thinking through local disaster risk management planning. *Procedia-Social and Behavioral Sciences,* 21:73-82.

Blochel, Kevin, Amanda Geniviva, Zachary Miller, Matthew Nadareski, Alexa Dengos, Emily Feeney, Alyssa Mathews, Jonathan Nelson, Jonathan Uihlein, Michael Floeser, Jörg Szarzynski, and Brian Tomaszewski, 2013, A Serious Game for Measuring Disaster Response Spatial Thinking. *ArcUser,* 16 (3):12-15.

Cheng, Z., F. Hao, Z. JianYou, and S. Yun, 2010, Research on design of serious game based on GIS. In: 2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design (CAIDCD), pp. 231-233.

Lee, Jongwon, and Robert Bednarz, 2012, Components of spatial thinking: Evidence from a spatial thinking ability test. *Journal of Geography,* 111 (1):15-26.

Michael, David R, and Sandra L Chen. 2005. *Serious games: Games that educate, train, and inform*: Muska & Lipman/Premier-Trade.

National Research Council. 2006. *Learning to Think Spatially: GIS as a Support System in the K-12 Cirriculum*. Edited by National Research Council. Washington, DC: The National Academies Press.

Ohmori, Nobuaki, Yasunori Muromachi, Noboru Harata, and Katsutoshi Ohta, 2003, Simulation Model for Activity Planning (SMAP): GIS-based gaming simulation. In: Selected Proceedings from the 9th World Conference on Transport Research. Elsevier, Oxfordpp.

Taylor, Holly A, and Thora Tenbrink, 2013, The spatial thinking of origami: evidence from think-aloud protocols. *Cognitive processing,* 14 (2):189-191.

United Nations University Institute for Environment and Human Security. *Education* n.d. [cited 28 April 2014]. Available from http://www.ehs.unu.edu/article/read/education.
US Department of Homeland Security. 2006. National Planning Scenarios.

# Type-2 Fuzzy Sets Applied to Geodemographic Classification

P.F. Fisher[1], N.J. Tate[1], A.Slingsby[2]

[1]Department of Geography, University of Leicester, Leicester LE1 7RH UK
Email: {pff1; njt9}@le.ac.uk

[2]Department of Computer Science, School of Informatics, City University, London EC1V 0HB UK,
Email: a.slingsby@city.ac.uk

## 1. Introduction

Fuzzy set theory has been widely and successfully applied to model uncertainty in a variety of geospatial contexts, including the classification of land cover (Foody 1996), and geodemographics (Grekousis and Hazichristos 2012). However, as noted by Fisher et al. (2007) such work predominantly makes use of fuzzy sets as originally specified by Zadeh (1965); characterised by crisp membership functions (Mendel and John, 2002). Zadeh (1975) extended his initial ideas to define fuzzy sets with fuzzy membership functions (Mendel and John 2002) and the nomenclature 'type-1' and 'type-2' is now used to refer to these different forms. Fisher and Tate (2014) employed type-1 fuzzy sets to soften a geodemographic classification (the UK Output Area Classification: OAC) of the City of Leicester UK. In this paper we extend that work to explore the use of type-2 fuzzy sets to the OAC.

## 2. Methods and data

Conventional geodemographics classifications are hard, assigning a single class to each output zone. However, in the process potentially useful information on variation in both geographic and classification spaces is effectively lost (Longley and Goodchild, 2008) and it is impossible to discriminate between zones which exhibit 'strong' and 'less strong' degrees of belonging to a class. This becomes more critical when exploring the local variation (e.g., by town or city) of a global (national) classification. While usually lost, this information may be very useful in applications (Slingsby et al., 2011). Fortunately various methods exist to soften hard classifications to derive multiclass memberships (Bezdek, 1981), especially fuzzy c-means (FCM) clustering (Bezdek, 1981) and possibilistic c-means (PCM). The latter replaces the constraint (Equation 1) of the FCM with a more inclusive constraint (Equation 2); in other words in FCM memberships for one particular case or object are constrained to sum to 1, but in PCM they are simply constrained to be in the range 0 to 1, the sum being no more than the number of classes.

$$\sum_{i=1}^{c} \mu_{ij} = 1, \quad \forall_j \in \{1, \ldots, n\} \tag{1}$$

$$0 < \sum_{i=1}^{c} \mu_{ij} \leq c, \quad \forall_j \in \{1, \ldots, n\} \tag{2}$$

At the core of both is an iterative process between class centroid calculation/update and distance-based membership calculation for $c$ classes and $n$ elements (zones) which optimises an objective function (Equation 3; Krishnapuram and Keller, 1993; Kruse et al. 2013; Pal et al, 2004):

$$J_{PCM}(\mathrm{X}, \mathrm{U}_{PCM}, \mathrm{C}) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} d_{ij}^{2} + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} \left(1 - \mu_{ij}\right)^{m} \quad (3)$$

where the possibility of zone $\mathbf{x}_j$ being in each class/cluster $c_i$ is given in Equation 4.

$$\mu_{ij} = \frac{1}{1 + \left(\frac{d_{ij}^{2}}{\eta_i}\right)^{\frac{1}{m-1}}} \quad (4)$$

Here $m$ is the fuzzifier and $d_{ij}$ the distances in classification space from each zone to the class centroid. $\eta_i$ is the distance of the "cross-over" point where $\mu_{ij}$ = 0.5 (Krishnapuram and Keller 1993), and is obtained from Equation 5.

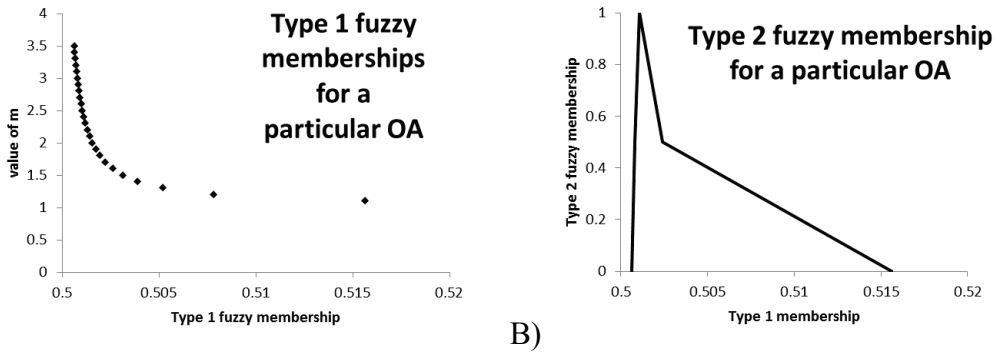
$$\eta_i = K \frac{\sum_{j=1}^{n} \mu_{ij}^{m} d_{ij}^{2}}{\sum_{j=1}^{n} \mu_{ij}^{m}} \quad (5)$$

The UK 2001 Output Area Classification (OAC) is a free geodemographic classification which at the highest taxonomic level is comprised of seven classes named 'Supergroups'. This classification employed a variant of hard c-means and critically in addition to the assigned class, distances $d_{ij}$ are also available for all *n* elements and *c* classes. Fisher and Tate (in press) employed [Equation 4] to create possibilistic memberships for each of the seven classes for each census reporting zone (Output Area - OA) for the City of Leicester. They compared PCM outcomes with fuzzy memberships from the equivalent FCM calculation favouring the PCM approach because of the constraint change in Equations 1 and 2. Following general practice (Bezdek 1981 among others) Fisher and Tate (in press) selected a crisp value of $m$ = 2. However, $m$ may have any value greater than 1, and following the method of Fisher (2010; see also Hwang and Rhee, 2007) by allowing $m$ to vary [1.1 to 3.5 in this instance] we can generate type-2 fuzzy sets for each Output Area.

## 3. Results

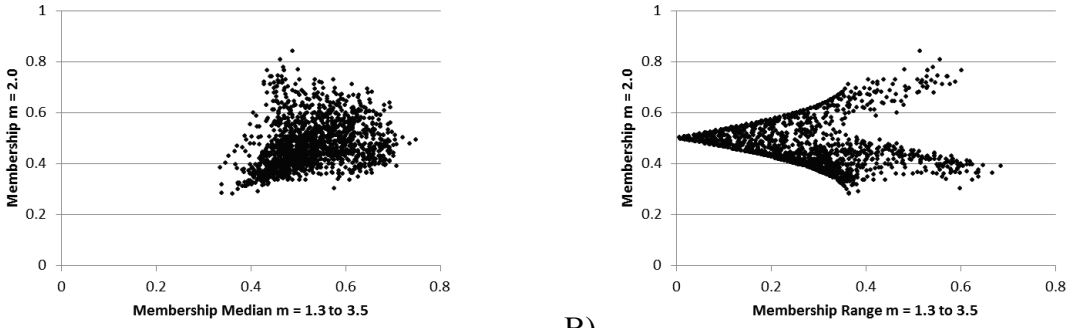
In Figure 1A, for one particular OA within Leicester, the distribution of all type-1 fuzzy memberships are shown plotted against the values of $m$ which yielded them. For convenience type-2 fuzzy sets can be summarised by taking appropriate summary statistics from the distribution of type-1 fuzzy memberships. Thus for the type-1 memberships for one particular OA shown in Figure 1A, the minimum, 1st, 2nd and 3rd quartiles, and maximum are used to summarise the type 2 fuzzy set in Figure 1B. These summary values are assigned memberships of 0, 0.5, 1, 0.5 and 0, respectively, in the type-2 fuzzy set (Figure 1B).

To establish that the type-2 memberships provide new information, the relationship to the type-1 memberships was examined. The relationship where $m$ = 2.0 for the City Living Supergroup is illustrated in Figure 2, and these membership values are seen to not be a good predictor of either the median or the range of the type-2 memberships. Furthermore, Figure 3 shows that neither the median nor the range are predictably related. Figures 2 and 3 are both for the City Living Supergroup only, but the same patterns of poor statistical prediction are repeated for all other Supergroups.




A) B)

**Figure 1.** For one particular Output Area and the City Living Supergroup A) shows the distribution of type 1 fuzzy memberships plotted against valuations of *m* which yielded them, and B) shows the form of the type 2 fuzzy set from the five summary statistics of that distribution discussed in the text.




A) B)

**Figure 2.** Scatterplots of membership range and membership median over m valuations from *m* = 1.1 to 3.5 (horizontal axis) against the type-1 membership (vertical axis) where *m* = 2.0 (the recommended valuation) for the City Living Supergroup

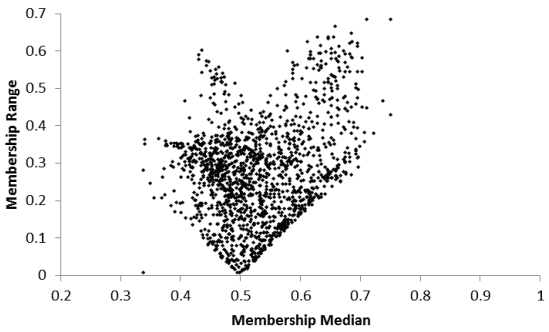



**Figure 3.** Scatterplot of membership range against membership median over the range of *m* valuations from *m* = 1.1 to 3.5 for the City Living Supergroup

Figure 4 shows four type 2 fuzzy sets for the City Living Supergroup derived as in Figure 1B, for those OAs with the largest and smallest median (2nd quartile), and the largest and smallest range (maximum – minimum) of membership values. The median shows the degree to which the OA is typical of the class; the OA with the largest being the most typical and the smallest the least. The range shows the degree to which a particular output area is a good example of that Supergroup. The OA with the smallest range is the most representative; all memberships for different valuations of *m* have similar values. The OA with the largest range is the poorest representative; the memberships are most varied. Shapes and

distributions of the seven graphs are remarkably similar showing that the range of memberships in each of the four types of OA are similar for all Supergroups.

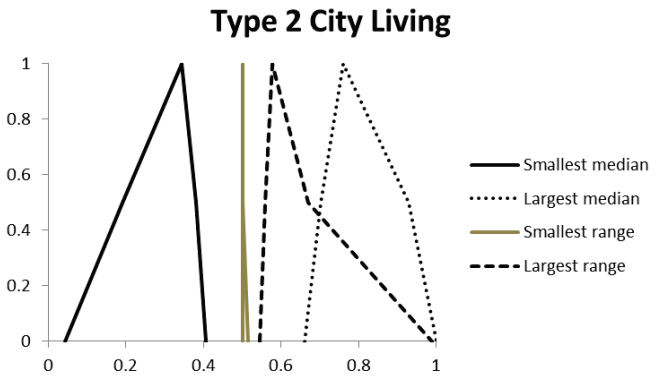



**Figure 4.** For the City Living Supergroup in the Output Area Classification the distributions of Type 2 fuzzy sets are shown for the four Output Areas having the smallest and largest range and median values.

## 4. Conclusion

Our research has revealed that for each Supergroup, we can subset OAs into distinct classes on the basis of whether they exhibit sensitivity to *m,* and to differentiate between those zones which display similar type-1 fuzzy memberships.

## References

Bezdek, JC, 1981, *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York.

Fisher, PF, 2010, Remote sensing of land cover classes as type 2 fuzzy sets. *Remote Sensing of Environment,* 114(2): 309-321.

Fisher PF, and Tate, NJ, in press, Modelling Class Uncertainty in the Geodemographic Output Area Classification, *Environment and Planning B*.

Foody GM, 1996, Fuzzy modelling of vegetation from remotely sensed imagery. *Ecological Modelling*, 17: 3-12.

Grekousis G and Hatzichristos T, 2012, Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods. *Applied Geography,* 34:125–136.

Hwang, C and Rhee, FC-H, 2007, Uncertain fuzzy clustering: Interval type-2 Fuzzy approach to c-means. *IEEE Transactions on Fuzzy Systems* **15**: 107-120.

Krishnapuram R and Keller JM 1993, A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems* **1:** 98–110.

Kruse R, Borgelt C, Klawonn F, Moewes C, Steinbrecher M, Held P, 2013, *Computational Intelligence: A Methodological Introduction* (Springer-Verlag, London).

Longley PA, Goodchild MF, 2008, The use of geodemographics to improve public service delivery. In: J Hartley, C Donaldson, C Skelcher, and M Wallace (eds), *Managing to Improve Public Services*. Cambridge University Press, Cambridge 176–194.

Mendel JM and John RI, 2002, Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems*, 10(2): 117-127.

Pal NR, Pal K, Keller JM, Bezdek JC, 2004, A new hybrid c-means clustering model in *Proceedings of the IEEE International Conference on Fuzzy Systems, Budapest,* (IEEE) pp 179–184.

Slingsby A, Dykes J. & Wood J, 2011, Exploring Uncertainty in Geodemographics with Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 17(12): 2545-2554.

Zadeh LA, 1965, Fuzzy sets. *Information and Control,* 8: 338–353.

Zadeh LA, 1975, The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences,* 8(3): 199-249.

# Does London need a separate geodemographic classification?

C.G. Gale[1], A.D. Singleton[2] and P. A. Longley[1]

[1]UCL Department of Geography, Gower Street, London, WC1E 6BT, UK.
Email: {chris.gale; p.longley}@ucl.ac.uk

[2]Department Geography and Planning, University of Liverpool, Liverpool L69 3GP, UK.
Email: alex.singleton@liverpool.ac.uk

## 1. Introduction

Geodemographic classifications are summary indicators of the social, economic, demographic and built characteristics of small neighbourhood areas. Typically classifications are built for the most disaggregate scale at which attributes of areas can be made available albeit that, as a technique, geodemographics can also be applied at more aggregate scales. Within the UK, there is a lineage of freely available small area geodemographic classifications that have been built from Census data outputs. The two most recent examples are the 2001 Output Area Classification (2001 OAC) and the 2011 Output Area Classification (2011 OAC), both of which are three-tiered hierarchical classifications of the UK. Output Areas (OA) are the smallest spatial element of UK Census geography and primary unit of dissemination for the last two UK Censuses.

## 2. 2001 OAC and 2001 LOAC

The top level of the 2001 OAC consists of seven clusters, or 'Supergroups'. Across the UK these Supergroups are relatively evenly distributed (see Table 1), this is not however the case when looking at only London. Table 1 shows that the 'Multicultural' Supergroup represents 56.1% of the population, and the 'City Living' Supergroup another 21.4%. Together these two groups represent over a quarter of London's population. This is caused by the high diversity of London being accommodated within a national classification; in important respects the UK is set apart from the prevailing characteristics of its capital city. Evidence would suggest therefore that London is a separate entity in geodemographic terms as is demonstrated by Longley et al. (2011). This indicates London would be better suited as being grouped together with other world cities - rather than as a region within the UK. Petersen et al. (2011: 177) suggest that "national classifications tend to be represented regionally by one dominant neighbourhood type" and that the remaining types then diminish exponentially.

Table 1. 2001 OAC Supergroup distributions.

| 2001 OAC Supergroups | UK OAs | London OAs |
|---|---|---|
| Blue Collar Communities | 16.1% (35,837) | 2.5% (606) |
| City Living | 7.5% (16,637) | 21.4% (5,174) |
| Countryside | 12.4% (27,681) | 0.1% (21) |
| Prospering Suburbs | 21.2% (47,250) | 7.4% (1,782) |
| Constrained by Circumstances | 14.9% (33,165) | 2.5% (592) |
| Typical Traits | 18.3% (40,769) | 10.1% (2,430) |
| Multicultural | 9.7% (21,721) | 56.1% (13,535) |

(Counts are in brackets)

In an attempt to resolve the issue of London and the 2001 OAC, Petersen et al. (2011) created a regional classification – the 2001 London Output Area Classification (2001 LOAC). This followed the same procedures outlined in Vickers and Rees (2007) when creating the 2001 OAC, but used a London only dataset – creating new cluster names and descriptions in the process. Conceived at the Supergroup level, the names and frequencies of each group are shown in Table 2. The rationale for the 2001 LOAC provokes wider questions of motivation, specification and estimation of freely available open geodemographics. For example, the motivation to create the 2001 LOAC was based upon the results of the 2001 OAC in London not being as representative as they could have been. Creating the 2001 LOAC resolved this particular issue, but as a consequence the ability to compare geodemographic characteristics with other areas in the UK was lost.

Table 2. 2001 LOAC Supergroup distributions.

| 2001 LOAC Supergroups | London OAs |
|---|---|
| Suburban | 10.4% (2,506) |
| Council Flats | 15.2% (3,678) |
| Asian Quarters | 11.3% (2,716) |
| Central District | 14.1% (3,409) |
| Blue Collar | 12.9% (3,114) |
| City Commuter | 14.7% (3,542) |
| London Terraces | 21.4% (5,175) |

(Counts are in brackets)

## 3. 2011 OAC and 2011 LOAC

The release of the 2011 UK Census and subsequent creation of the 2011 OAC provided an opportunity to address how London is represented within a national level geodemographic classification. Although the 2011 OAC bears methodological similarities to its predecessor the number of clusters formed to create the three-tiered hierarchy differed. In total eight clusters, also termed Supergroups, formed the top level of the hierarchy. The addition of an extra cluster meant the distribution of the Supergroups across the UK, as identified in Table 3, was not as uniform when compared to the 2001 OAC. Nevertheless, the 2011 OAC can be considered to provide a clear and easy way of interpreting the socio-demographics of the UK. Table 3 also shows how the addition of an eighth cluster impacted London. No single cluster represents more than 37% of London's OAs, an improvement compared to the 2001 OAC. This is however tempered by the fact three Supergroups now represent 85% of London's OAs.

Table 3. 2011 OAC Supergroup distributions.

| 2011 OAC Supergroups | UK OAs | London OAs |
|---|---|---|
| Rural Residents | 11.75% (27,300) | 0.1% (15) |
| Cosmopolitans | 5.65% (13,125) | 14.3% (3,584) |
| Ethnicity Central | 5.10% (11,849) | 37.0% (9,263) |
| Multicultural Metropolitans | 10.12% (23,502) | 32.9% (8,233) |
| Urbanites | 16.66% (38,697) | 9.1% (2,285) |
| Suburbanites | 20.17% (46,850) | 4.6% (1,141) |
| Constrained City Dwellers | 11.68% (27,135) | 1.1% (281) |
| Hard-Pressed Living | 18.87% (43,838) | 1.0% (251) |

(Counts are in brackets)

The 2011 OAC can be considered to provide a better representation of London compared to its predecessor. This can in part be explained by results from the 2011 UK Census showing the rest of England and Wales is becoming more ethnically diverse like London (ONS, 2012). This would indicate London is becoming less unique within England and Wales, thereby becoming less of an outlier in a national geodemographic classification. This is however a slow process, meaning London is still likely to exhibit characteristics not found elsewhere in the UK for the foreseeable future. As such, the improvements in London's representation offered by the 2011 OAC were not considered satisfactory enough for some organisations. An example of this is the Greater London Authority (GLA) who only require a geodemographic perspective of London. This led to the GLA commissioning the creation of the 2011 London Output Area Classification (2011 LOAC) to provide the level of detail required for use in London only applications (shown in Table 4).

Table 4. 2011 LOAC Supergroup distributions.

| 2011 LOAC Supergroups | London OAs |
|---|---|
| Intermediate Lifestyles | 12.9% (3.241) |
| High Density and High Rise Flats | 12.5% (3,138) |
| Settled Asians | 11.6% (2,913) |
| Urban Elites | 9.5% (2,376) |
| City Vibe | 14.1% (3,523) |
| London Life-Cycle | 12.9% (3,224) |
| Multi-Ethnic Suburbs | 14.8% (3,709) |
| Ageing City Fringe | 11.7% (2,929) |

(Counts are in brackets)

The 2011 LOAC, similar to its own predecessor, was based on the same methodology as the national classification with a London only dataset. The names of the Supergroups and their distribution across London's OAs are shown in Table 4. Compared to the 2011 OAC, the eight Supergroups provide a more even distribution across London. Figure 1 shows how representative the 2011 LOAC Supergroups are of the 2011 OAC Supergroups. The 'Settled Asians' Supergroup for example consists almost entirely of the 2011 OAC Supergroup 'Multicultural Metropolitans'; in contrast the 'Ageing City Fringe' Supergroup covers multiple 2011 OAC Supergroups. The compositions of most 2011 LOAC clusters do however differ significantly from those of the 2011 OAC in London. This, and their more even distribution, would suggest the 2011 LOAC Supergroups provide a more comprehensive overview of London's urban socio-economic structure when compared to their 2011 OAC counterparts.

## 4. Applications for a London Geodemographic Classification

The 2011 LOAC provides more options for London focused applications. An example of this is coding London's Metropolitan Police Service's (MPS) Public Attitude Survey (PAS) by geodemographic type. The PAS is a rolling survey that quantifies the satisfaction with local policing across London. A sample of the PAS contains over 125,000 unique geo-coded records collected between April 2006 and September 2013. Coding these responses to the 2011 OAC reveals that 69.9% of responses have been collected from areas assigned to the 'Multicultural Metropolitans' and 'Ethnicity Central' Supergroups. In contrast 1% of responses have come from the 'Hard-Pressed Living' Supergroup. The same analysis performed using the 2011 LOAC reveals a much smaller disparity, ranging from 14.8% to 9.5% between Supergroups.

The enhanced representation offered by the 2011 LOAC allows for further analysis to be performed on the dataset, such as multinomial logistic regression (Moutinho and Hutcheson,

2007) or log linear analysis to predict the probabilities of each geodemographic type having particular attitudes towards the police. Such analysis performed using the 2011 OAC would be complicated by the small sample sizes for several of its Supergroups. As such, the creation of the 2011 LOAC provides a potentially valuable tool to the MPS to better understand how their activities are perceived in London from a geodemographic perspective.

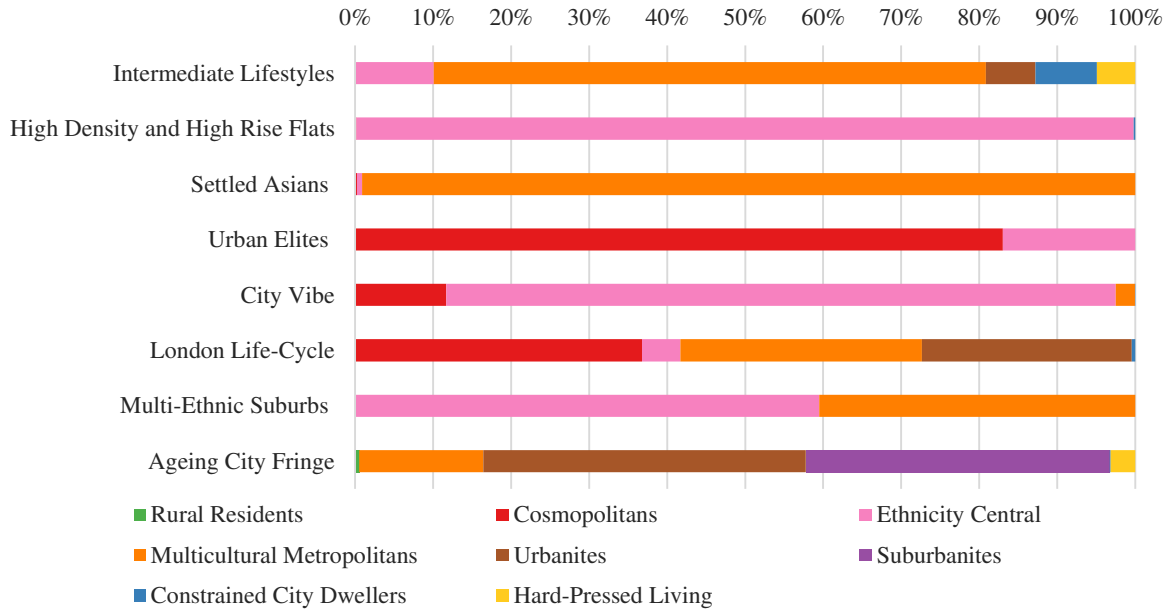



Figure 1. How representative the 2011 LOAC Supergroups are of the 2011 OAC Supergroups.

## 5. Conclusion

The unique nature of London makes the creation of its own classification essential to best understand the geodemographic characteristics of its resident population. The creation of classifications like the 2011 LOAC allow a greater number of London focussed applications to incorporate a geodemographic perspective. An example of this is the PAS, which by incorporating the 2011 LOAC with its own results allows for more in-depth analysis and prediction to be performed. Applications like this would be made more challenging using a national level classification, highlighting the continued need for a London focussed solution.

## Acknowledgements

This work is part of the project - Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks (www.ucl.ac.uk/cpc), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1). The data provided by Metropolitan Police Service (London) is highly appreciated.

## References

Longley, P. A., Cheshire, J. A. and Mateos, P. (2011) 'Creating a regional geography of Britain through the spatial analysis of surnames', Geoforum, 42, pp. 506–516.

Moutinho, L. and Hutcheson, G. D. (2007) 'Store choice and patronage: a predictive modelling approach', International Journal of Business Innovation and Research, 1(3), pp. 233–252.

ONS (2012) 'Ethnicity and National Identity in England and Wales 2011', Office for National Statistics, [online] Available from: http://www.ons.gov.uk/ons/dcp171776_290558.pdf (Accessed 20 August 2013).

Petersen, J., Gibin, M., Longley, P., Mateos, P., Atkinson, P. and Ashby, D. (2011) 'Geodemographics as a tool for targeting neighbourhoods in public health campaigns', Journal of Geographical Systems, 13(2), pp. 173–192.

Vickers, D. W. and Rees, P. H. (2007) 'Creating the UK National Statistics 2001 output area classification', Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(2), pp. 379–403.

# Spatio-temporal demographic classification of the Twitter users

Paul Longley, Muhammad Adnan, Guy Lansley
University College London, Department of Geography, Gower Street, London, WC1E 6BT.
Email: plongley@geog.ucl.ac.uk ; m.adnan@ucl.ac.uk

## 1. Introduction

Use of social media continues to increase day by day, with implications for the creation of 'big' data – Twitter alone was forecast to have created 1.8 zettabytes of data in 2011. Users of the likes of Twitter, Facebook, Flickr, LinkedIn, Bebo, and Orkut are frequently mobile users of the developing range of smartphones and tablet devices. The availability of such data has profound implications for the geodemographic analysis of human settlement structure. Hitherto, small area measures of neighbourhood conditions have been based only upon the night-time socioeconomic characteristics of residential areas and selected physical characteristics of their built environments (Martin et al, 2012). Although useful for guiding resource allocation decisions for many private and public goods (Longley, 2005), these static and essentially cross sectional views provide only limited insights into the functioning of settlement systems and the temporal heterogeneity that characterises their component parts. The advent of new data sources derived from social media also has profound implications for our understanding of behaviour in virtual as well as observable space, and interactions between the two. Taken together, the prospect of developing composite cyber-geodemographic measures offers the prospect of better understanding the dynamic as well as the static organisation of human settlements (Pahl 1970).

This paper presents an initial work towards the creation of geo-temporal geodemgoraphic classifications by using the Twitter social media data. London was chosen as the study area because of its high incidence of users and the consequent expectation that higher penetration might be associated with lower demographic bias. Our research has focused on the classification of twitter users as geodemographic applications draw upon the characteristics of the population living in small neighbourhood areas. Our analysis uses given and family names of the social media users to determine their ethnicity and age. The ethnicity and age attributes are used as the input to the clustering analysis for the creation of demographic classifications of the twitter users at different temporal scales.

## 2. The Geography of the Twitter Users

The Twitter Streaming API (Twitter, 2012) can be used to download a 1% sample of the geotagged tweets. For this paper, the Twitter Streaming API was used to download geo-tagged Tweets for London during the period 11$^{th}$ September, 2012 to 28$^{th}$ February, 2013. The fields downloaded from the API included the user name, latitude and longitude from which the Tweet was sent, time and tweet message content. A total of 4.1 million (4,103,072) geo-tagged Tweets were downloaded, sent by a total of 230,972 unique users.

## 3. Ethnicity and Age Analysis of the Twitter Users

While registering their on Twitter, users are required to enter their name or other identifying data in the 'User Name' field. In many cases, tokens other than given and family names are entered, as in 'MysticMIND', 'JustinBieber Home' etc. Our text analytic work suggests that approximately 65% of Tweets are identified by recognisable forename-surname pairs. The 'User Name' field was divided into separate 'forename' and 'surname' fields for these users, as illustrated in Table 1.

| User Name | Forename | Surname |
|---|---|---|
| Kevin Hodge | Kevin | Hodge |
| Jose De Franco | Jose | De Franco |
| Carolina Thomas, Dr | Carolina | Thomas |
| CunninghamMichaelDr | Michael | Cunningham |
| Coles, Ian Stewart | Ian | Coles |

Table 1: 'User Name' field divided into separate 'forename' and 'surname' fields

Our text analytics detected probable forename-surname pairs for c.2.6 million (2,660,433) of the 4.1 million Tweets. These 2.6 million Tweets were sent by 158,375 unique users. In the next step, Onomap (Mateos et al, 2011) was used to assign forename and surname pairs to predicted ethnic groups. Approximately the forename-surname pairs of the 2.1 million (2,064,633) Tweets, sent by 122, 107 unique users, were successfully classified. Individuals were assigned to a total of 67 ethnic groups. For creating a geo-temporal demographic classification, the following 8 ethnicity variables were created from the 67 onomap ethnic groups.

V1: British & Irish
V2: West Europeans
V3: East Europeans
V4: Greek & Turkish
V5: South East Asians
V7: Chinese
V8: Other

For the age analysis, given names (forenames) were used in order to estimate the likely ages of their bearers, using an enhanced version of CACI's (London, UK ; http://www.caci.co.uk/) Monica system. This uses c. 7 million records drawn from consumer dynamics files to identify the frequencies of 11,700 different given names within five year age bands. The nature of the source data means that younger adult cohorts are under-represented relative to the UK population as a whole, and there are no records pertaining to individuals below the age of 18 at all (despite their making up 22% of the UK population).

In order to mitigate this bias, all names from birth certificates with frequencies of two or above, representing 9.7 million individuals, were acquired from the Office of National Statistics for the years 1994 – 2011. (It was not, however, possible to identify a source that would have enabled the names of international immigrants under the age of 18 to be identified.) The birth certificate data were disaggregated into five year age group bands for consistency with the Monica classification, and both sources were reweighted to fit the age distribution of the 2011 Census of Population for England and Wales. The average ages of bearers of the names in the resulting dataset ranged from just 2 years to 83 years of age. Using the above datasets, for each twitter user, the following age variables were created:

V9: Age: 10 – 20 years
V10: Age: 21 – 35 years
V11: Age: 36 – 50 years
V12: Age: 51 – 65 years
V13: Age: 65 plus years

## 4. Computing the spatio-temporal demographic classifications

For computing the demographic classifications at different time scale, the k-means cluster analysis (Vickers & Rees, 2007) was performed on the 13 demographic variables. Prior to the

classification process, the variables were aggregated to the 633 CAS wards in the Greater London. The cluster analysis was performed at the following three temporal scales.

Week Days: 7.01 a.m. to 6.00 p.m. during the week days
Week Evenings: 6.01 p.m. to 12.00 a.m. during the week days
Week Nights: 12.01 a.m. to 7.00 a.m. during the week days

Following figures (1-3) show the cluster analysis results for different temporal scales.

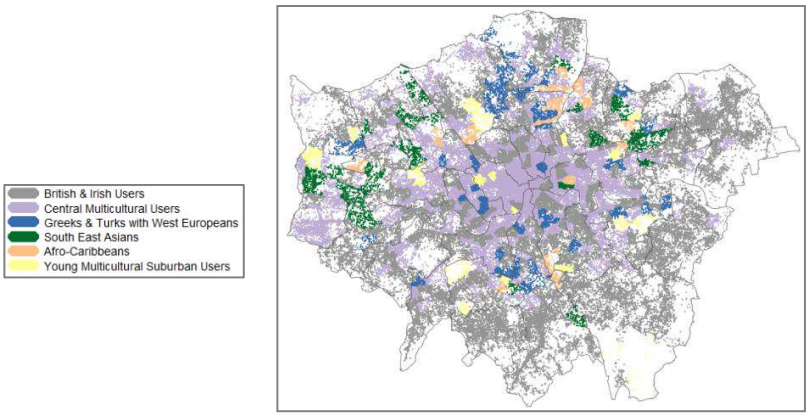



Figure 1: Spatio-temporal demographic classification during the week days

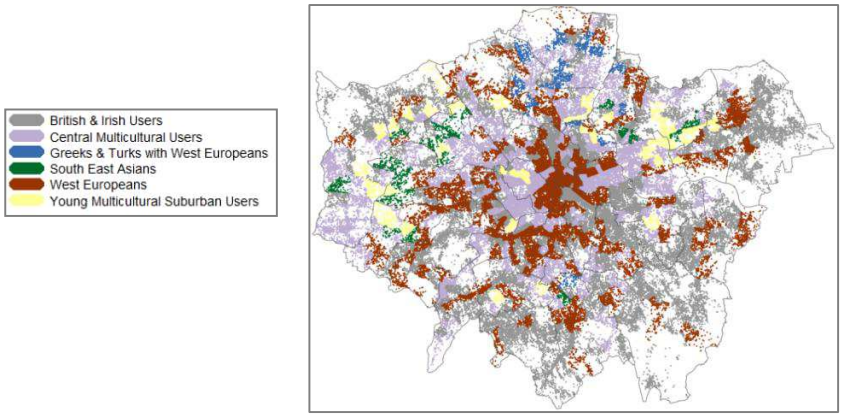



Figure 2: Spatio-temporal demographic classification during the week evenings

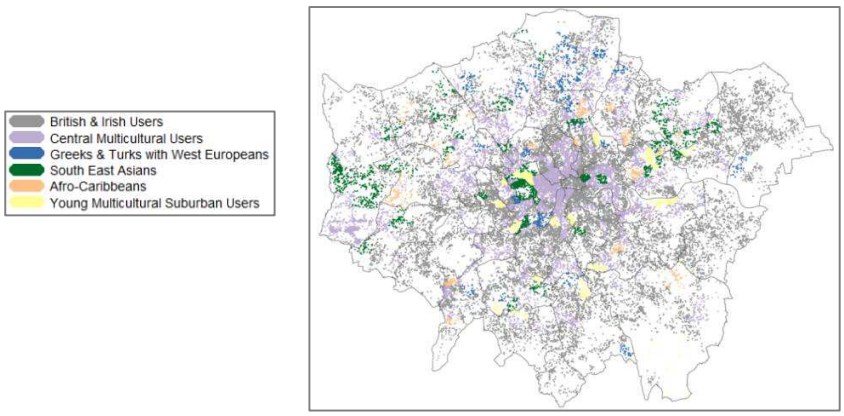



Figure 3: Spatio-temporal demographic classification during the week nights

Cluster analysis detected spatial variations in the activity patterns of the twitter users. During the week days, the center of the city is occupied by the multicultural users represented by the 'Central Multicultural Users' cluster. 'British & Irish Users' is the largest cluster of the classification, which is an indicator of the high tweeting activity by the native users in the city. Areas of the high tweeting activities of different ethnic minority groups in London were also identified. These areas are represented by the clusters 'South East Asians', 'Afro-Caribbeans', and 'Greeks and Turks with West Europeans'.

For the week evenings, cluster analysis detected a different cluster called 'West Europeans'. This cluster is indicative of the areas where young West Europeans twitter users have high tweeting activity during the evenings of the week days.

Tweeting activity during the week nights is very low (Figure 3). Cluster analysis, performed on the week nights data, detected the similar clusters as for the week days. However, there are spatial differences in the activity patterns of different clusters between week days and week nights.

## 7. Conclusion

This paper has presented a preliminary work towards the creation of geo-temporal demographic classifications of the social media users. The three geo-temporal demographic classifications provide an insight into the spatial activity patterns of the different kind of social media users in the Greater London. Future work will extend this analysis to other temporal scales i.e. weekend. Future research will also emphasis on the use of additional demographic attributes of the social media users in the classification process.

## Acknowledgements

This work was completed as part of the EPSRC research Grant "The Uncertainty of Identity: Linking Spatiotemporal Information in the Real and Virtual Worlds" (EP/J005266/1).

## References

Longley, P. A. 2005. "A renaissance of geodemographics for public service delivery". Progress in Human Geography, 29: 57-63

Martin, D., Cockings. S., Harfoot., A. 2012. "Development of a geographical framework for Census workplace data". Journal of the Royal Statistical Society: Series A (Statistics in Society) (In Press).

Mateos, P., Longley, P. A., O'Sullivan. D. 2011. "Ethnicity and population structure in personal naming networks". PLoS ONE (Public Library of Science), 6(9) e22943, 1-12

Pahl, R. 1970. Whose City? Penguin, Harmondsworth

Twitter (2012). "What is Twitter ?". From https://business.twitter.com/basics/what-is-twitter/.

Vickers, D.W., Rees, P.H. (2007). Creating the National Statistics 2001 Output Area classification. Journal of the Royal Statistical Society, Series A. 170(2), 379-403.

# A method for calculating regional differences in land cover / land use change and error

A.J. Comber[1], P. Fisher[1], H. Balzter[1], S. Johnson[1], B. Ogutu[1], B. Cole[1]

[1]Department of Geography, Centre for Climate and Landscape Research, University of Leicester, Leicester, LE1 7RH, UK
Email: {ajc36; pff1; hb91; sj239; boo7, bc132}@le.ac.uk

## 1. Introduction

The correspondence matrix is the classic framework in remote sensing of land cover for reporting error and change. It is variously referred to as an error matrix, a confusion matrix or a change matrix depending on the context of its use. The measures that are commonly generated include Type I and Type II errors in validation exercises and loss and gain in change analyses. As the correspondence matrix describes the counts of pixels or objects allocated to different classes by model and observation (validation) or at Time 1 and Time 2 (change), the diagonal, off-diagonal, row totals and column totals can be used to generate probabilities of, for example class to class errors or transitions. In this way the correspondence matrix can be considered as a visual representation of a series of logistic regression models. In geographical analyses researchers are often interested in comparing the results of some analysis in one region with those found in another. The regions may describe different underlying environmental processes (e.g. geological, climatic, etc.) or may relate to different anthropogenic activities related to ownership or management. In remote sensing, this is often done by constructing separate correspondence matrices for each region under consideration, and then comparing statistics derived from each matrix. For example, by comparing the probabilities that Class A is mapped as Class B or that Class A has changed to Class B, derived from correspondence matrices constructed using data from different region. This paper suggests an alternative method for statistically comparing regions and for evaluating change based on logistic regression.

## 2. Data

Consider the two land cover or land use maps in Figure 1. In 1965 the National Trust commissioned a land use survey of the coastline in England, Wales and Northern Ireland, which has recently been scanned to digital format. The survey is in the process of being updated using visual interpretation of high-resolution imagery. There were 14 classes in the original survey (note that classes 7, 12 and 13 have been excluded from the change analysis). The mapped areas, covering a roughly 20km by 20km region, indicate the pattern of land around the ports of Felixstowe and Harwich in the East of England. It is obvious from Figure 1 that this area has experienced a considerable amount of land use change, especially within and around the port areas, but also in the agricultural the surrounding agricultural areas (yellow). Ongoing work will update the entire coastline of England, Wales and Northern Ireland, defined by the geographic extent of the remit of the National Trust.

## 3. Evaluating change

Changes in land use and land cover are typically summarised using a change matrix and a number of standard measures can be generated such as per class loss and gain rates and their parallels in a validation matrix, user and producer accuracies. Many of the measures that are commonly derived from correspondence matrices (whether related to change or validation)

describe probabilities, for example in the case of validation, of Class A in the modelled data being Class A in the reference data (producer accuracy), or in the case of Class A at Time 1 not being the same class at Time 2 (change probability). Many of these row and column indices can be described by logistic regression of the data summarised by the contingency table and Comber (2013) provides a full treatment of the relationship between logistic regression, probability and the correspondence matrix.

1965 2014

Figure 1. The 1965 and 2014 land use data.

However, in many situations there is considerable interest in understanding how land use changes or errors relate to different underlying factors. In the case of change this could be related to land ownership, local government policy for building developments and land management. In the case of error, changes could be to do with bio-geographical zone, landscape type or regions related to the used remote sensing imagery. The usual approach is to compare the probabilities arising from separate correspondence matrices, calculated from data covering each area. The main caveat in seeking to explore the differences in change or error between regions is that when regions are considered in this way, the probabilities are specific to each correspondence matrix, to each regional analysis. They cannot be easily related. The typical statements can take the form: *The probability that areas mapped as Class 9 in the Region 1 analysis will have changed when mapped in 2014 is 0.25.* Table 1 shows correspondence constructed from the 1965 and 2014 data.

## 4. Proportionality

It is possible to develop generalised linear models to estimate likelihoods of change as a function of the regions using a model that expects proportionally equal levels of change in each zone. The counts of pixels that have and have not changed were summed for the different zones in a table of counts. In this, the rows indicated whether change had occurred or not and the columns indicated the zone. To test for an association, $A$, between the row and column effects, the Poisson regression model was applied:

$$A(c_{ij}) = \log(r + C_i + R_j) \qquad \text{(Equation 1)}$$

where the count in column $I$ and row $j$ is denoted by $c_{ij}$ and has a Poisson distribution, r is an intercept term, $C_i$ is a column effect and $R_j$ is a row effect, is compared against the model:

$$A(c_{ij}) = \log(r + C_i + R_j + I_{ij}) \qquad \text{(Equation 2)}$$

where the extra term $I_{ij}$ is an interaction effect between rows and columns. If this is significantly zero, then it suggest that there is some degree of association between the row

and column effects. In the context land cover / land use change detection, this approach can be used to test for association between zones and change. The counts were cross-tabulated for the different zones (Table 2) to show the losses in different zones.

Table 1. Correspondence matrices, in hectares, for 3 regions (rows 1965, columns 2014).

| R1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 9 | 10 | 11 | 14 | Same | Loss | Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 15.8 | | | 50.3 | | | | 9.6 | 0.1 | | | | 0.21 | 0.79 | 0.29 |
| **1** | | 172.8 | | | | 0.8 | | 1.2 | | | | | 0.99 | 0.01 | 0.40 |
| **2** | 2.3 | | | | | | | | | | | | 0 | 1 | 0 |
| **3** | | | | 3.2 | | | | | | | | | 1 | 0 | 0.97 |
| **4** | | | | 2.0 | | | | | | | | | 0 | 1 | 0 |
| **5** | | | | | | 11.6 | | | | | | | 1 | 0 | 0.44 |
| **6** | | | | | | | | | | | | | 0 | 0 | 0 |
| **8** | 2.4 | | | 0.1 | | | | 24.6 | | | | | 0.91 | 0.09 | 0.58 |
| **9** | 1.6 | 93.4 | | 48.4 | | 8.2 | | 15.6 | 642.1 | 37.2 | 4.5 | | 0.75 | 0.25 | 0 |
| **10** | | | | | | | | | | 56.7 | | | 1 | 0 | 0.40 |
| **11** | | 19.6 | | 6.1 | | 0.2 | | 7.7 | | | 54.4 | | 0.62 | 0.38 | 0.08 |
| **14** | | | | | | | | | | | | | 0 | 0 | 0 |

| R2 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 9 | 10 | 11 | 14 | Same | Loss | Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 9.7 | | | | | | | | | | | 1.4 | 0.87 | 0.13 | 0.12 |
| **1** | | 37.1 | | | | | | | | | 0.1 | | 1 | 0 | 0.42 |
| **2** | | | 0.4 | | | | | | | | | | 1 | 0 | 0 |
| **3** | | | | | | | | | | | | | 0 | 0 | 1 |
| **4** | | | | | 0.3 | | | | | | | | 1 | 0 | 0.98 |
| **5** | | | | | | | | | | | | | 0 | 0 | 0 |
| **6** | | | | 1.2 | 17.6 | | | | | | | | 0 | 1 | 0 |
| **8** | | | | | | | | | | | | | 0 | 0 | 0 |
| **9** | 1.3 | 23.2 | | | | | | | 1168.6 | 1.7 | | 7.1 | 0.97 | 0.03 | 0 |
| **10** | | 0.3 | | | | | | | | 75.4 | | 0.2 | 0.99 | 0.01 | 0.04 |
| **11** | | 3.8 | | | | | | | | 1.3 | 65.9 | 0.5 | 0.92 | 0.08 | 0 |
| **14** | | | | | | | | | | | | | 0 | 0 | 1 |

| R3 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 9 | 10 | 11 | 14 | Same | Loss | Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 12.5 | | | | | | | 99.5 | | | 0.9 | | 0.11 | 0.89 | 0.68 |
| **1** | | 86.5 | | | | 1.1 | | | | | | | 0.99 | 0.01 | 0.24 |
| **2** | 1.2 | | | | | | | | | | | | 0 | 1 | 0 |
| **3** | | | | 2.9 | | | | 3.6 | | | | | 0.44 | 0.56 | 0.98 |
| **4** | | | | | | | | | | | | | 0 | 0 | 0 |
| **5** | | | | | | 3.6 | | | | | | | 0.99 | 0.01 | 0.82 |
| **6** | 2.8 | | | | | 5.8 | 6.1 | 35.2 | | | 27.9 | | 0.08 | 0.92 | 0 |
| **8** | | | | 0.8 | | | | 2.9 | | | | | 0.79 | 0.21 | 0.99 |
| **9** | 22.9 | 28.0 | | 158.8 | | 9.0 | | 87.5 | 826.5 | 30.8 | 5.2 | 11.5 | 0.70 | 0.30 | 0 |
| **10** | | | | | | | | | 3.7 | 55.6 | | | 0.94 | 0.06 | 0.36 |
| **11** | | | | | | 0.8 | | | | | 7 | | 0.99 | 0.01 | 0.33 |
| **14** | | | | | | | | | | | | | 0 | 0 | 1 |

Table 2. Regional change in 2014, with the regions sorted right to left in decreasing area..

| Region 3 | Region 2 | Region 1 |
|---|---|---|
| No Change | 10665487 | 13574621 |
| Change | 5371146 | 597259 |

Values of $I_{ij}$ were estimated by fitting Equation 2 to the regional data and the resulting coefficients were related to a comparative index of change for each of the row categories, using the formula:

$$CHANGE = 100(\exp(I_{ij}) - 1) \qquad \text{(Equation 3)}$$

Due to the way the interaction terms are calibrated, this compares each column *category j* (zones) against a 'reference' category. A value of 0 suggests the likelihood of access for *category j* is the same as for the reference category. A value of +50 for *category j* suggests access is one-and-a-half times as likely as the reference category, a value of -50 that it is half as likely, and so on. The reference category is the zone with the largest area. For each of the coefficients, *CHANGE* was calculated and the results are shown in Table 3.

Table 3. The change likelihoods for different zones, relative the largest zone (Region 3).

| Region | Change likelihood | Pr(>\|z\|) |
|---|---|---|
| Region 2 | -91.26 | 0.000 |
| Region 1 | -36.95 | 0.000 |

It is possible to consider how the likelihoods of change differ between the regions and the classes. Table 4 shows the changes occurring to Class 1 in different regions and the results of the likelihood analysis are shown in Table 5. These indicate that Class 1 changes (loss in this case) are 3% more likely in Region 3 than in Region 1 and 86% less likely in Region 2 than in Region 1.

Table 4. The area of Class 1 in different regions found to have changed.

| | Region 3 | Region 2 | Region 1 |
|---|---|---|---|
| No Change | 1727870 | 864704 | 370783 |
| Loss | 21038 | 10831 | 636 |

Table 5. The change likelihood from Class 1, relative the largest zone (Region 1).

| Region | Change likelihood | Pr(>\|z\|) |
|---|---|---|
| Region 3 | 2.9 | 0.017 |
| Region 2 | -85.9 | 0.000 |

## 4. Discussion

There a whole host of issues that could be discussed here, but in brief, likelihood statistics are relatively easy to compute from either the raw data or from the correspondence matrix. Cross tabulation of how different factors interact (regions and classes) can be applied to a generalized linear model, and is applied to predict the frequency of occurrence of the count under a Poisson distribution. The exponentials allow likelihoods to be generated. Such statistics are also widely applicable, especially in remote sensing and GIS analyses where such models have the capacity to generate more informative reporting of change and error than simple consideration of different correspondence matrices.

## 4. Acknowledgement

This work was supported by the National Trust, Neptune Coastal Land Use Mapping Project.

## References

Comber AJ, (2013), Geographically weighted methods for estimating local surfaces of overall, user and producer accuracies. *Remote Sensing Letters*, 4(4): 373-380.

# Assessing metrics of user quality in a simple land cover validation task

C. F. Salk[1], T. Sturn[1], S. Fritz[1], L. See[1]

[1]International Institute for Applied Systems Analysis
Ecosystem Services and Management Program
Schlossplatz 1
Laxenburg A-2361
Austria
Email: salk@iiasa.ac.at

## 1. Introduction

Crowdsourcing (Howe, 2006), or Volunteered Geographic Information (VGI) (Goodchild, 2007) is a cost effective method of completing mapping tasks such as generating up to date street maps (e.g. via OpenStreetMap; Neis et al., 2012), for post-disaster mapping (Zook et al., 2010), or validating land cover (Clark and Aide, 2011; Fritz et al., 2012) that would otherwise be prohibitively expensive. However, the quality of volunteer labour is highly variable, requiring methods to assess individual volunteers' ability to complete tasks accurately (Goodchild and Li, 2012). Because checking each volunteer-completed task would negate any benefit of crowdsourcing, it is necessary to develop statistical methods to generate probabilities that tasks are completed rigorously with no, or very limited, reference data. In this paper we use a simple VGI task, identification of cropland in satellite images and ground-based photographs using a game called Cropland Capture to assess several metrics of user quality.

## 2. Methods

Cropland Capture is a game with worldwide coverage in which users classify imagery either from satellites or ground-based photographs (See et al., 2014). For each image, a user has three classification choices: 'cropland', 'not cropland' and 'maybe cropland'.

For each user, we computed several performance measures which form the basis for these analyses. User-specific output quantity was measured with the total number of ratings performed and the number of images receiving a non-maybe (i.e. cropland or non-cropland) rating. The quality of user output was assessed in several ways. Each user's rate of agreement with majority-based classifications was calculated, omitting responses of 'maybe' and images evenly split between cropland and non-cropland. For images rated more than once by a user, a self-contradiction rate was computed as the proportion of subsequent ratings agreeing with initial ratings. The proportion of 'maybe' ratings was computed as a metric of user caution. User bias was assessed by calculating the proportion of cropland and non-cropland ratings that were correct.

We used regression analyses to test the relationship between different aspects of users' performance. For certain variable pairs, we have reasons to hypothesize a causal relationship, for instance, the hypothesis that accuracy increases with user experience. Because the independent variable (user experience, measured as the number of points rated) is measured without error, it is appropriate to use standard ordinary least squares (OLS) regression in this case. However, for other variable pairs, there is no theoretical reason to expect most user metrics to be more accurately measured than others as we believe that all such metrics are reflections of an underlying (and unobserved) user quality variable. In these

cases, we employ major axis regression (also known as type II regression) as implemented in the R package lmodel2. This method does not assume any underlying differences between the variables being analysed, and unlike OLS regression, returns the same result when the identity of the variables is switched.

In some cases we were interested to uncover not just patterns among typical users, but also patterns among top users as they contribute disproportionately to the eventual land cover classification goals of Cropland Capture. For this purpose we used quantile regressions, implemented in the R package lmodel2. Quantile regression has the added benefit of not relying on the distributional assumptions of OLS regression so is particularly suited to the heteroskedastic patterns seen between many of the user quality variables. However, quantile regression still assumes error only in the dependent variable, so is best suited to relationships where one variable is precisely known.

## 3. Results

A total of 2,361 different users contributed ratings to Cropland Capture between mid-November 2013 and mid-January, 2014. These users provided ratings on 96,038 images for a grand total of 2,547,843 user-image rating combinations. This total includes a small percentage of images that were seen more than once by a particular user to test repeatability of their ratings. The number of images rated by individual users is log-normally distributed and ranges from 2 to 201,831 images. Overall, users disagreed with the crowd classification of points 6.5 % of the time. Crowd-contradiction rate for users with more than 1000 ratings ranged from 1.1 % to 17.1 %. When users rate a point more than once, they agree with their initial rating 96.1% of the time.

Users who were more likely to self-contradict were also more likely to give ratings in disagreement with the crowd. This pattern holds regardless of whether based on raw ($p<.0001$; $R2=.345$) or log-transformed variables ($p<.0001$; $R2=.361$). Users' use of the 'maybe' classification decreased with self-contradiction rate (Figure 1). This decreasing trend was seen regardless of whether we used the raw variables ($p<.0001$, $R2=.0657$) or log-transformations of the variables with values of 0 omitted ($p<.0001$, $R2=.0671$).

User experience, as measured by the log of the number of points rated, excluding responses of 'maybe', shows a complex relationship with rating quality. Users' median rate of disagreement with the crowd decreases significantly with increasing experience in the game (quantile regression with $\tau=.5$; $p<.0001$, $\rho=.77$). However, among top performing users, this relationship was reversed (quantile regression with $\tau=.1$; $p<.0001$, $\rho=.53$; Figure 3). When only users with substantial experience (>1000 images classified) were considered, these relationships weakened considerably. Median users still showed an improved crowd agreement with experience ($p<.0055$, $\rho=.04$), but the slope of 10th percentile users became non-significantly negative ($p=.6984$).

User self-contradiction rate showed similar patterns to crowd-contradiction rate with increasing user experience. Because users received only occasional repeat images, this calculation was only possible for users who performed large numbers of ratings. In contrast to crowd-contradiction rate, the median user's self-contradiction rate increased slightly with experience ($p=.0111$, $\rho=.04$). The top users (10th percentile of self-contradiction rate) showed a more strongly positive trend toward making more errors with increasing experience ($p<.0001$, $\rho=.013$). Only the bottom users (90th percentile of self-contradiction rate) showed a decrease in self error rate with experience ($p=.0123$, $\rho=.02$).

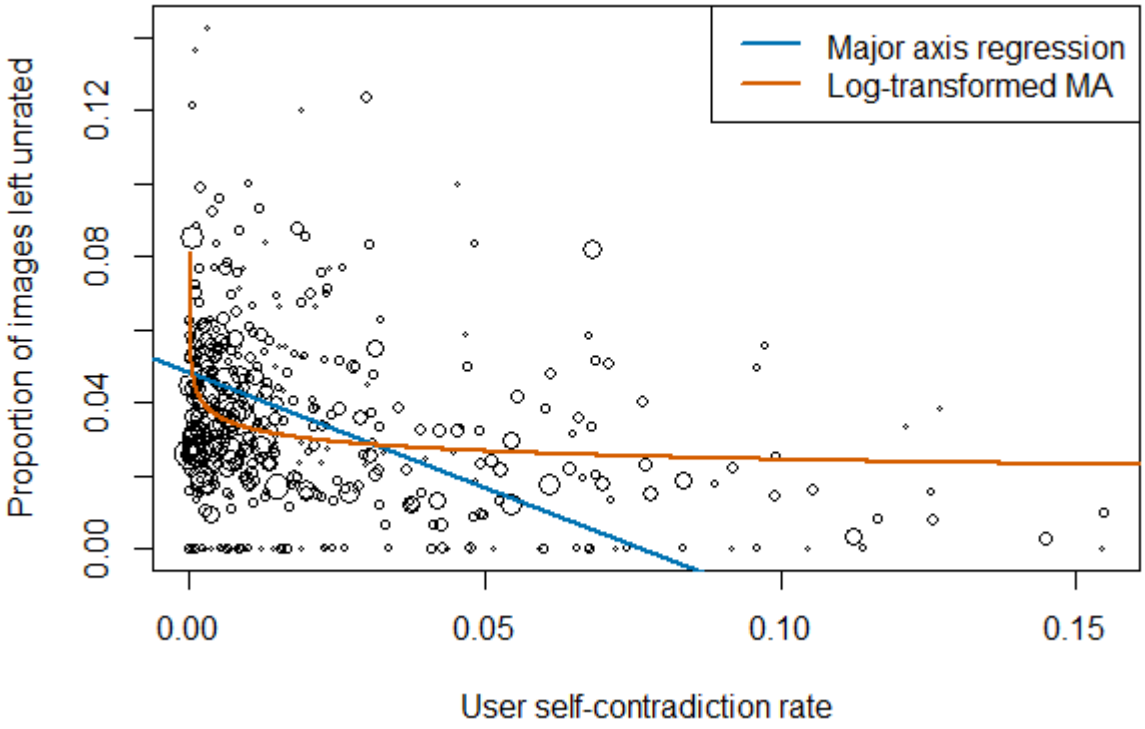



**Figure 1**. The relationship between users' admission of uncertainty about presence of cropland in an image as a function of their rate of giving self-contradictory results in multiple classifications of the same image in the Cropland Capture game. Each point corresponds to a single user. Only users who have classified more than 1000 images are included in this figure. The straight line is a major axis regression which treats variables equally, rather than assuming all error is in the dependent variable. The curved line is a major axis regression on log transformations of the same variables (including only non-zero values). Circle size is proportional to number of images rated by a user.

## 4. Discussion

Our results suggest that volunteers are highly effective at rating photographs and satellite imagery for the presence of cropland. Consistency was high with regards both to individual users' previous ratings of the same images, and with other users' ratings of those images. That user self-contradiction rate and crowd-contradiction rate are positively correlated suggests that self-contradiction rate is a useful method of evaluating the quality of contributed data.

While these findings are based on analysis of large post-game datasets, they suggest certain applications to the eventual goal of run-time evaluation of contributors' quality. Many of the metrics reported on here can be regularly recalculated and provided as feedback to users, either cumulatively or for a window of recently rated points. Such run-time evaluation would have direct application as feedback to help users improve their performance or as part of a ranking system in which users receive credit or payment based on the quality and quantity of ratings they provide.

## Acknowledgements

This work was funded by the CrowdLand project (ERC grant #617754).

## References

Clark, M.L., Aide, T.M., 2011, Virtual Interpretation of Earth Web-Interface Tool (VIEW-IT) for collecting land-use/land-cover reference data. Remote Sensing 3:601–620.

Goodchild, M.F., 2007, Citizens as sensors: the world of volunteered geography. GeoJournal 69:211–221.

Goodchild, M.F., Li, L., 2012, Assuring the quality of volunteered geographic information. Spatial Statistics 1:110–120.

Howe, J., 2006, The rise of crowdsourcing. Wired Magazine, 14.06.

Neis, P., Zielstra, D., Zipf, A., 2012, The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Future Internet 4: 1-21. doi:10.3390/fi4010001

See, L., Sturn, T., Perger, C., Fritz, S., McCallum, I., and Salk, C., 2014, Cropland Capture: A gaming approach to improve global land cover. AGILE 2014, Castellon Spain, 3-6 June 2014.

Zook, M., Graham, M., Shelton, T. and Gorman, S. 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. World Medical & Health Policy, 2:7–33.

# Sensitivity analysis of Support Vector Machine land use change modelling method

Mileva Samardžić-Petrović[1], Branislav Bajat[1], Miloš Kovačević[1], Suzana Dragićević[2]

[1] Faculty of Civil Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia
Email: {mimas; bajat; milos}@grf.bg.ac.rs

[2] Spatial Analysis and Modeling Laboratory, Department of Geography, Simon Fraser University, 8888 University Dr., Burnaby, B.C., V5A1S6, Canada
Email: suzanad@sfu.ca

## 1. Introduction

When using machine learning (ML) approaches for land use (LU) change modelling the main goal is to find a function that is the best approximation of the nonlinear problem that represents the complex LU change process. Support Vector Machines (SVM) is one ML method capable of solving nonlinear problems and has been applied to various disciplines such as ecology (Drake et al. 2006), hydrology (Tripathi et al. 2006) and remote sensing (Brown et al. 2000). Interest in using the SVM method for LU changes modelling has grown in recent years (Yang et al. 2008, Okwuashi et al. 2012). However, as a relatively new method, SVM is insufficiently researched in LU change modeling, particularly in relation to its sensitivity to parameter changes, attribute selection, data sampling choices, and data representation.

The main objective of this research study is to conduct a sensitivity analysis investigation on a SVM-based LU change model with respect to attribute selection and parameter changes. The efficient application of ML methods, including SVM, requires the selection of appropriate attributes (features). Attribute selection is an important stage of modelling as some attributes can have small or no predictive power at all, and hence "confuse" the ML process. Various methods for attribute selection exist. However, this study uses Info Gain (IG), Gain Ratio (GR) and Correlation-based Feature Subset (CFS) (Witten et al. 2011). Moreover, the efficient application of SVM requires the selection of an optimal combination of parameters. Because the Radial Basis Function (RBF) (Abe, 2005) was used as the kernel function for SVM, model sensitivity was analyzed for changes to two parameters; $\gamma$ of the RBF and penalty C. Method have

## 2. Experiment

### 2.1 Study area and data representation

The study area included the Zemun Municipality, part of the territory of the City of Belgrade, Republic of Serbia. In 2013, the administrative area was approximately 17.5km x 8km and included the old city and two suburban areas.

The data used for this study included: three orthophoto images for the years 2003, 2007 and 2011; maps of actual LU classes obtained from the Urban Planning Institute of Belgrade, and; publicly available population census data (Statistical Office of the Republic of Serbia).

The study area was represented as a rectangular cell grid of 10m spatial resolution. Also, each grid cell was represented as an n-dimensional real vector $\mathbf{x}^t$ ($\mathbf{x}^t=<x^t_1, x^t_2,\ldots, x^t_n>$), where

coordinate $x^t_i$ represents the value of the i-th attribute (corresponding to LU class and created urban attributes) associated with the cell x, at a particular time t. The various urban attributes considered were: Euclidian distance ed to municipality centre, city centre (old core of Belgrade), Danube and Sava rivers, green areas, railway, highway, main road, streets of category I and II, and Population Change Index (PCI) between two censuses for the years 2002 and 2011. PCI provides a standardized measure for comparing population changes over time (Bajat et al. 2013).

Since the goal of the SVM model is to predict future land use changes, it is necessary to prepare a minimum of two datasets to represent the study area at three different moments in time (t-1, t and t+1). In order to build the model $\mathbf{x}^{t-1} \rightarrow y^t$, SVM uses training dataset $(\mathbf{x}^{t-1}, y^t)$, which contains $\mathbf{x}^{t-1} = <x^{t-1}_1, x^{t-1}_2, \ldots, x^{t-1}_n>$ as input attributes and $y^t$ as the output attribute to be predicted. For this study, $y^t$ contains nine LU classes: agricultural, wetlands, traffic areas, infrastructure, residential, commercial, industry, special use and green areas.

Based on the developed SVM model, $\mathbf{x}^{t-1} \rightarrow y^t$, and by using $\mathbf{x}^t = <x^t_1, x^t_2, \ldots, x^t_n>$ as input attributes, $y^{t+1}_p$ can be predicted. Therefore, the second dataset is a test dataset $(\mathbf{x}^t, y^{t+1})$ and it is used for independent validation of the built SVM model, achieved by comparing the predicted $(y^{t+1}_p)$ and real $(y^{t+1})$ LU classes at time t+1. The training dataset was created based on data from years 2003 and 2007, and the test data created from years 2007 and 2011.

The Kappa statistics was used to compare the model output with the real land use map for year 2011. The overall land use change for the study time period (2003-2011) was small (3%) relative to the overall study area size. Hence, in order to obtain more informative datasets it was necessary to conduct data sampling. Consequently, in this study the training and test datasets were created to contain the same number of changed and unchanged cells, thereby ensuring that all LU classes are proportionally represented.

## 2.2 Attribute Selection

From datasets S containing all of the created attributes, three subsets of attributes were selected using the IG, GR and CFS methods, and the results shown in Figure 1. The IG and GR methods rank the attributes independently of each other based on their measure of association with the LU class in time t, while the CFS method automatically determines a subset of k relevant attributes that are highly correlated with the LU class but uncorrelated with each other.

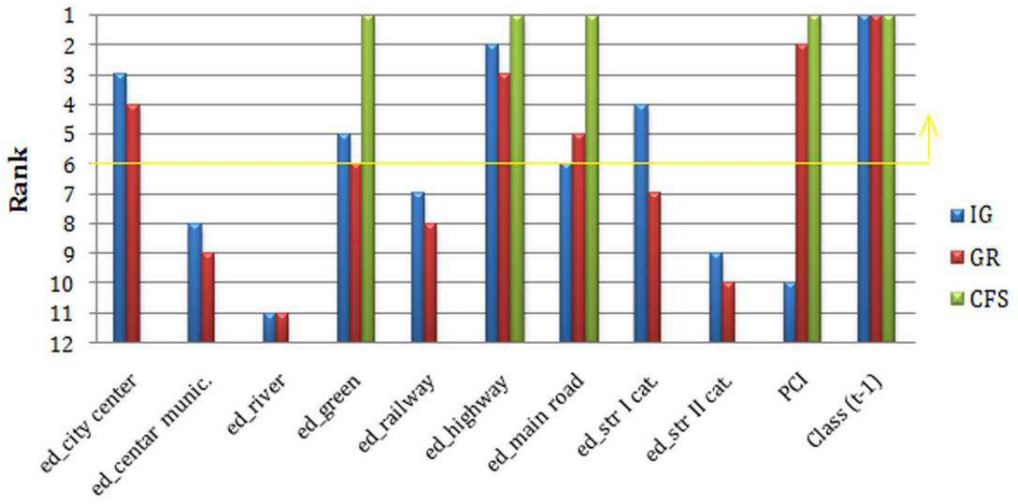



Figure 1. Comparisons of the attribute selection by IG, GR and CFS methods.

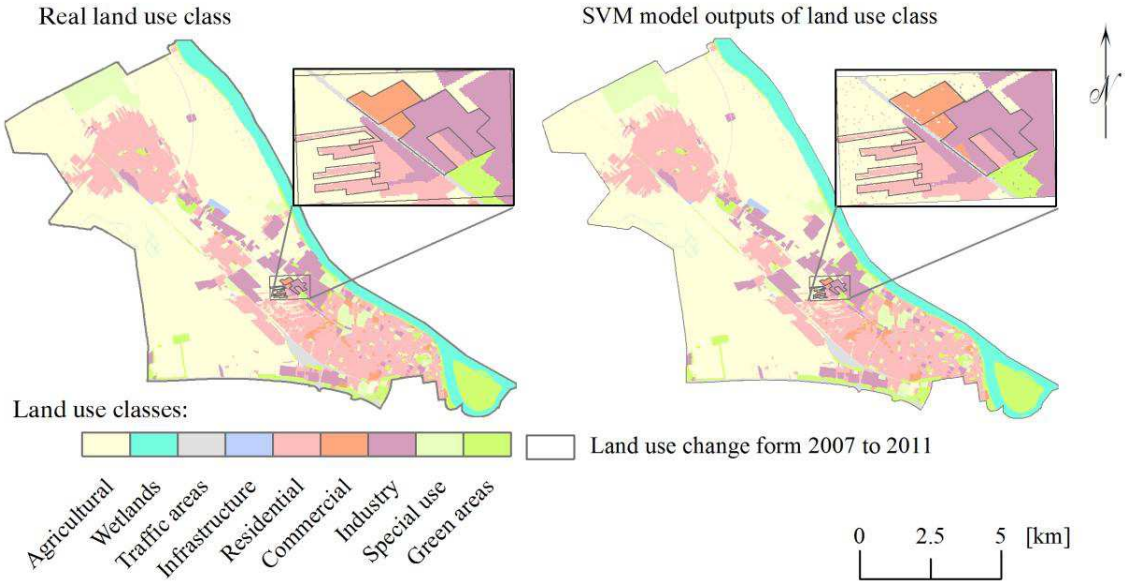



Figure 2. Generated land use changes with SVM model for 2011.

One of the obtained SVM model outputs for the generated land uses for year 2011 is shown in Figure 2. Moreover, the results indicate the CFS method selected subsets of five attributes. Therefore, in order to compare the sensitivity of models built with attributes selected using the three methods with respect to the SVM parameters, the five highest ranked attributes by IG and GR were selected and three datasets were created. Each dataset $S^{CFS}$, $S^{IG}$ and $S^{GR}$ contains training and test datasets for five selected attributes respectively for the CFS, IG and GR methods.

## 2.2 Sensitivity of model on SVM parameters and selected attributes

In order to demonstrate the significance of the attribute selection, additional models were created based on the datasets S. Values of validation measure for all models, built by using various SVM parameters and four different data representations are shown in Figure 3.

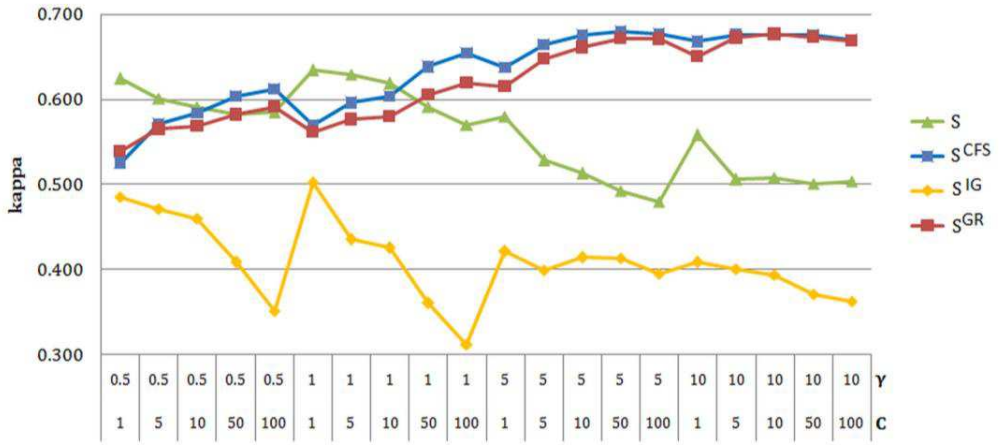



Figure 3. Kappa values for LU change models based on different values for SVM parameters γ and C for S, $S^{CFS}$, $S^{IG}$ and $S^{GR}$ datasets.

The results indicate that the $S^{CFS}$ and $S^{GR}$ datasets exhibit better kappa performance and are more robust to different SVM parameter combinations. Models built based on $S^{CFS}$ datasets are slightly better than the ones based on $S^{GR}$. Using S and $S^{IG}$, the LU model have less capability to predict changes and can be overfitted with higher values of parameters.

## 3. Conclusion

The obtained results indicate the subset of the same number of attributes selected by the CFS and GR methods increased kappa values, while attributes selected by the IG method decreased kappa values comparing to models built using all attributes. Using selected attributes by the CFS and GR methods resulted in a simple (less attributes – less complicated) model but with better performance and with less possibility to be overfitted with higher values of parameters. For the datasets used, the subset of k attributes selected by the CFS method provided slightly better models compared to the k highest ranked attributes by GR, and significantly better models compared to k highest ranked attributes by IG. In order to further explore sensitivity analysis of the SVM modelling approach, the future work will include using datasets covering different study areas and with different land use change dynamics. Furthermore, the k number of selected attributes was not investigated as the optimal number of attributes selected by IG and GR, which will also be pursued in future research.

## Acknowledgements

This study was supported by an award from the Ministry of Education, Science and Technological Development of the Republic of Serbia (Project No. III 47014) to the first three authors, and by a Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grant awarded to the fourth author.

## References

Abe S, 2010, Support vector machines for pattern classification, 2nd ed: Springer.

Bajat B, Krunić N, Samardzić-Petrović M and Kilibarda M, 2013, Dasymetric modelling of population dynamics in urban areas, Geodetski vestnik, 2013(57): 777–792.

Brown M, Lewis HG and Gunn SR, 2000, Linear spectral mixture models and support vector machines for remote sensing, Ieee Transactions on Geoscience and Remote Sensing 38 (5):2346-2360.

Drake JM, Randin C and Guisan A, 2006, Modelling ecological niches with support vector machines, Journal of Applied Ecology 43 (3):424-432.

Okwuashi O, McConchie J, Nwilo P, Isong M, Eyoh A, Nwanekezie O, Eyo E and Ekpo AD, 2012, Predicting future land use change using support vector machine based GIS cellular automata: a case of Lagos, Nigeria. Journal of Sustainable Development 5 (5):132-139.

Tripathi S, Srinivas VV and Nanjundiah R S, 2006, Dowinscaling of precipitation for climate change scenarios: A support vector machine approach, Journal of Hydrology 330 (3-4):621-640.

Witten IH, Frank E, and Hall MA, 2011, Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. 3rd ed: Elsevier.

Yang Q, Li X, and Shi X, 2008, Cellular automata for simulating land use changes based on support vector machines, Computers & Geosciences 34 (6):592-602.

# Scale Effects in Relating Movement to Geographic Context

C. Gschwend[1], P. Laube[2]

[1]Department of Geography, University of Zurich, Zurich, Switzerland
Email: christian.gschwend@geo.uzh.ch

[2]Institute of Natural Resource Sciences, Zurich University of Applied Sciences, Wädenswil, Switzerland
Email: patrick.laube@zhaw.ch

## 1. Introduction

Relating movement to context allows a better understanding of movement as traces of behavior, since movement is typically influenced by external factors such as the geographic context (Nathan et al. 2008). Within GIScience, movement analysis is mainly concerned with algorithmically detecting shape, arrangement, or interaction patterns on a geometric basis. Much less work has been done on relating movement to its embedding geographic context. Hence, this paper specifically addresses context-aware movement analysis, especially the little understood scale effects in quantifying the relation of movement to its context. Although there is previous scale-related movement research in GIScience (Laube and Purves 2011) as well as in behavioral ecology (Börger et al. 2006), most studies focus on one dimension of scale, be it spatial, temporal, or thematic scale, and do not consider interdependencies between the different scale dimensions.

To this end, we are specifically interested in revealing interdependencies between different dimensions of scales. So, we address the following research questions:

- How sensitive is the computation of a quantitative relation between movement and its embedding context to a systematic variation of the temporal, spatial and thematic analysis scales?
- When such scale sensitivity exists, can interdependencies between the different scale dimensions be identified and quantified?

An empirical study with movement data of chamois and terrain aspect as geographic context is carried out, in order to tackle these research questions.

## 2. Data and Methodology

In this study, GPS movement data of seven chamois (Table 1) from the Swiss National Park (in the South-East of Switzerland with an area of around 170 square kilometres) is related to the aspect of the terrain, in order to assess in what aspect classes the animals move, whilst the temporal scale of the movement data and the spatial and thematic scales of the aspect are systematically varied.

Table 1. Specifics for GPS movement data of chamois

| Parameter | Chamois |
|---|---|
| Time span | 12/2002 – 04/2010 |
| Mean time span per animal | 1.4 years |
| Temporal sampling rate | 10min (every 2nd Wednesday) / 4h |
| No. of animals | 7 (6 female, 1 male) |
| No. GPS points | 29'571 |
| Source | Swiss National Park |

The experiments relating movement to aspect are carried out at three spatial scales (4, 20, and 100 meters raster resolution), three thematic scales of the context (5, 9, and 17 aspect classes), and three temporal scales of the movement (10min / 4h, 30min / 12h, and 1h / 1 day). Furthermore, movement is modelled in two different ways in order to discover scale effects depending on the chosen conceptual movement model. First, movement is represented as the mere GPS points. As a second movement model, the Brownian Bridge Movement Model (BBMM) is realized, which represents movement in form of a probability density surface as a raster (Horne et al. 2007). So, movement and the terrain aspect are related based on the GPS fixes and the BBMM within the 99% volume contours. The point-based movement model is related to the terrain parameter by considering the aspect values in the exact location of the GPS fixes. In the case of the BBMM, each cell of the BBMM-raster is related to the nearest neighbor in the raster representing the terrain parameter and the probability density value is used to weight the aspect. Relative distributions for different combinations of spatial, temporal and thematic scales are statistically analyzed by assessing the differences quantitatively using the coefficient of variation.

## 3. Results

We present our results in Figure 1 that show how the relative distribution of the context variable aspect varies with different temporal scales of the movement, different spatial and thematic scales of the context, and different movement-context relation-methods. The different rows in Figure 1 represent the variation of the thematic scale of the aspect (5, 9, and 17 categories). Moreover, this figure consists of four columns including different scales for 'time' and 'space', a 'method' and a 'coefficient of variation' column. The 'time' column illustrates the relative distribution of aspect for three temporal scales (10min/4h, 30min/12h, 1h/1d; $t_1/t_2$: temporal scale of $t_1$ every 2nd Wednesday, else temporal scale of $t_2$) when the spatial scale is kept constant at 4 meters (@4m). The 'space' column demonstrates the effects of systematically varied spatial scales (4m, 20m, 100m) on aspect at a fixed temporal scale of 10min/4h (@10min/4h). The 'method' column illustrates the relative distributions of aspect for the two different methods used to relate movement models to aspect ('map pin' vs. 'BBMM-based'), using a point-based (GPS fixes) or a raster-based (BBMM) movement model. In the last column to the right, we present the 'coefficient of variation', where the bars reflect the within-class variation of the first three columns 'time', 'space' and 'method'. In the colour version, the top orange bar references the variation of the 'time' column, the green bar the 'space' column and the blue bar the 'method' column. We illustrate this in the dashed box in Figure 1, where the 'space' and 'method' columns vary more than the bars in the 'time' column. This is mirrored in the corresponding coefficients of variation (for 'space' and 'method': around 0.2, for 'time': 0.004).

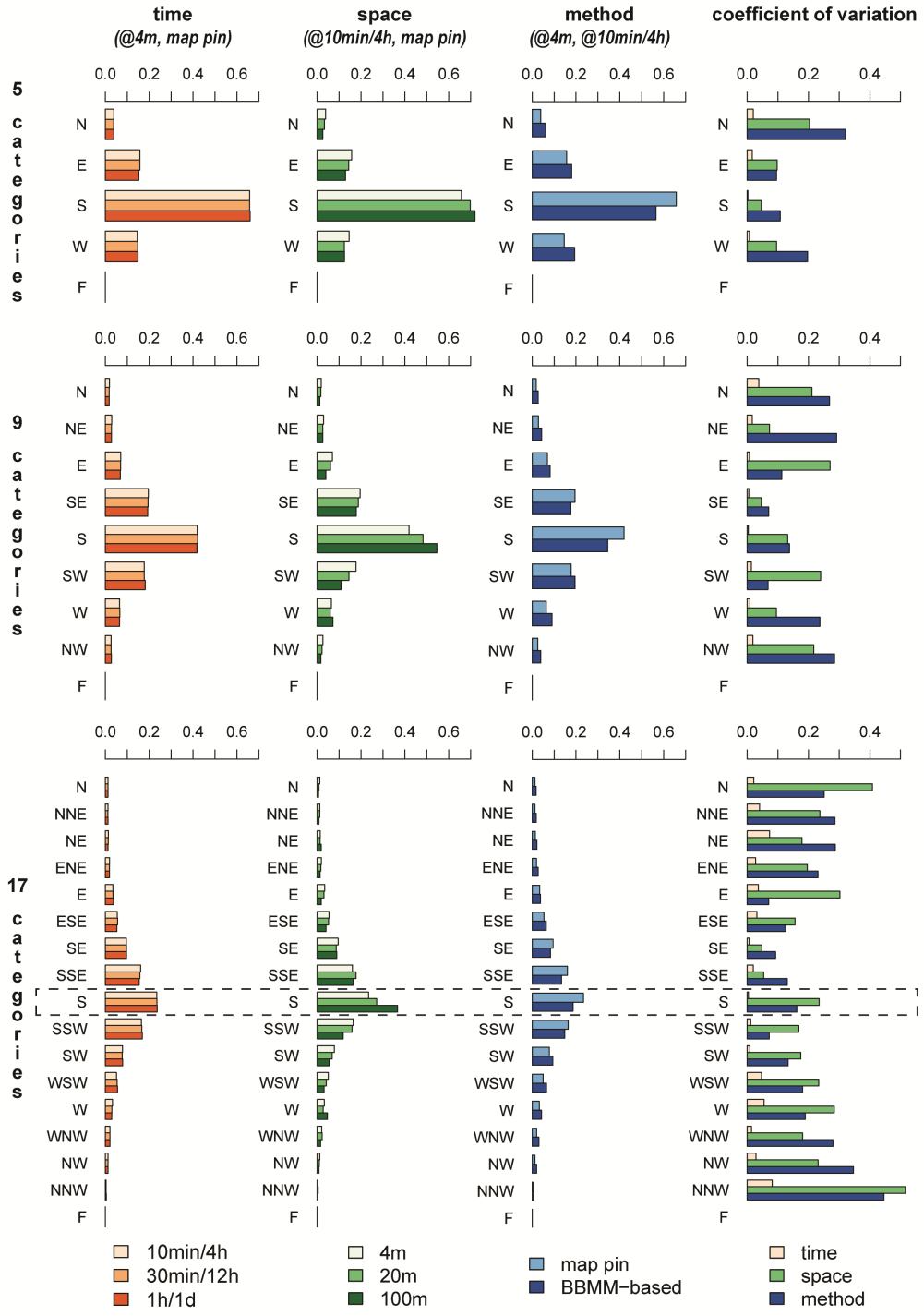



Figure 1. Relative distributions (0-1) of geographic context (terrain aspect) in relation to chamois' movement. Systematic variation of temporal scale of movement ('time' column), spatial and thematic scale of context ('space' column / y-axis), and relation-method ('method' column). 'coefficient of variation' (0-∞) column presents within-category variation.

### 3.1 Differences due to Scales and Methods

Considerable variations can be found within categories when varying scales and relation-methods. We sum up our most important findings in Figure 1 as follows:

- Variations due to spatial scale are larger (differences of up to 13%, coefficient of variation around 0.2) than variation due to temporal scale (differences negligible, coefficient of variation around 0.02).
- Coefficients of variation due to relation-methods are comparable to the ones caused by different spatial scales.
- With 17 categories of aspect, those categories with higher relative values of the distribution (e.g. South, dashed box), show in general smaller coefficients of variation than categories with a smaller share (e.g. North).

### 3.2 Interdependencies between Scales

Figure 2 shows coefficients of variation, which are computed in analogy to the procedure applied for the variation values in Figure 1 ('coefficient of variation' column). However, Figure 2a shows the coefficients of variation resulting from a systematic variation of the temporal scale (10min/4h, 30min/12h, 1h/1d, 'temporal scale effects') on *all* the spatial scales of the geographic context (4m, 20m and 100m). Similarly, Figure 2b shows variations caused by systematically varying spatial scale (4m, 20m, 100m, 'spatial scale effects') for *all* the temporal scales of the movement (10min/4h, 30min/12h and 1h/1d). For example, the smallest black bar in the dashed box in Figure 2a (17 categories, North) shows for a '4m' spatial scale a coefficient of variation of 0.02 when varying the temporal scale (10min/4h, 30min/12h, 1h/1d). The first result in the following list confirms the expectations with respect to the number of categories. However, Figure 2 reveals interesting statements about interdependencies between different sorts of scales:

- As expected, the more categories of aspect are introduced, the smaller relative values per category get, and the higher grow potential coefficients of variation (law of large numbers).
- Differences due to temporal scale get more pronounced with coarser spatial scale (Figure 2a).
- Spatial scale effects are more stable with regard to different temporal scales (Figure 2b).

Scale effects that are discussed in this paper, similarly arise for the 'map pin' as well as for the 'BBMM-based' relation-method. So in this case study, the movement-context relation-method has no effect on the revealed scale effects.

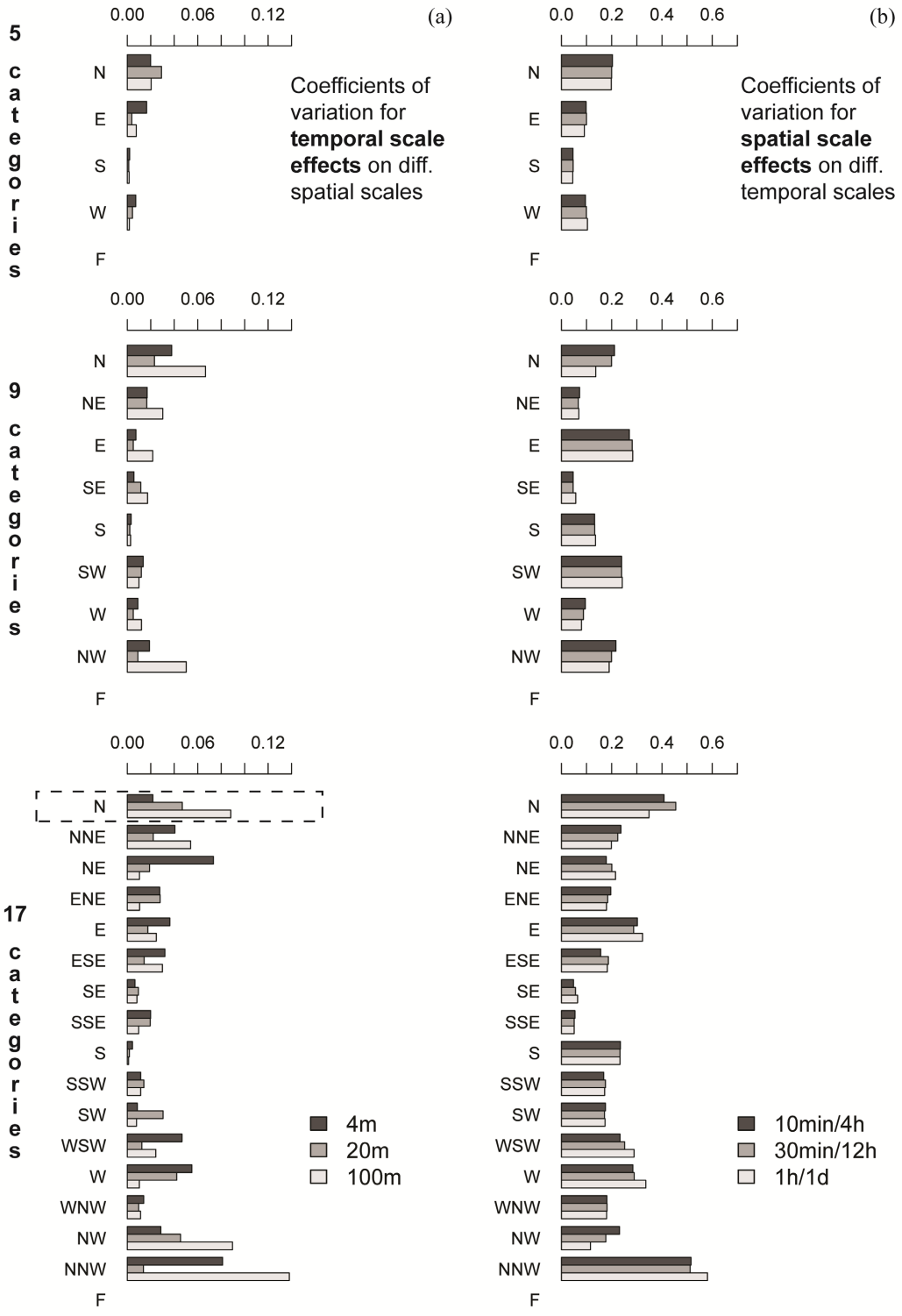



Figure 2. Coefficients of variation representing within-class variation due to variation of temporal (a) and spatial (b) scale across different thematic scales.

## 4. Discussion and Conclusions

In this paper, we illustrated with an empirical study the complex interplay between different types of analysis scales when relating movement to its embedding geographic context. In our study relating ungulate movement to terrain aspect the movement-context relation proved to be sensitive to some factors (spatial scale, thematic scale, relation-method), but not so much to others (temporal scale). Differences in preferences with regard to terrain's aspect might be more pronounced on a seasonal temporal scale for chamois. In terms of interdependencies between scales, our study suggests that the sensitivity of the results to the movement sampling rate depends on the spatial and thematic scale of the context. Similarly, the spatial granularity of the embedding context in turn matters more or less depending on the sampling rate of the movement. The key contribution of our work lies in providing quantitative evidence for the otherwise often overlooked complex interplay between the major scale dimensions in movement analysis.

## Acknowledgements

The authors thank Ruedi Haller and Flurin Filli, Swiss National Park, for providing the movement data of ungulates and the fine-grained Digital Elevation Model. This work was supported by the Swiss National Science Foundation, project 200021_129963.

## References

Börger L, Franconi N, Ferretti F, Meschi F, De Michele G, Gantz A and Coulson T, 2006, An integrated approach to identify spatiotemporal and individual-level determinants of animal home range size. *The American Naturalist*, 168:471-485.

Horne JS, Garton EO, Krone SM and Lewis JS, 2007, Analyzing animal movements using Brownian bridges. *Ecology*, 88:2354–2363.

Laube P and Purves RS, 2011, How fast is a cow? Cross-Scale Analysis of Movement Data. *Transactions in GIS*, 15:401-418.

Nathan R, Getz WM, Revilla E, Holyoak M, Kadmon R, Saltz D and Smouse PE, 2008, A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences*, 105:19052-19059.

# Routing through a continuous field constrained by a network

N. Karrais, A. Keler, S. Timpf
Department of Geography, University of Augsburg,
Alter Postweg 118, 86159 Augsburg, Germany
Email: {nicole.karrais; andreas.keler; sabine.timpf}@geo.uni-augsburg.de

## 1. Finding routes with least exposure

A person suffering from a chronic respiratory disease has to take current air quality into account for each errand that takes the person outside their home (Gehring et al. 2013). These individuals are interested in estimating ahead of time how high the exposure to e.g., particulate matter would be when leaving home and taking a specific route. In order to make a good decision, a selection of routes with minimal exposure to air pollutants should be offered, e.g. as a location-based service. This type of problem calls for mapping a changing continuous field onto a network in order to enable optimization for routing.

From a GIScience perspective the combination of surface (or volume) information with network information for the purpose of routing is an interesting problem. We leave aside the issue of how to sample such changing phenomena and assume that spatio-temporal surfaces have been provided either in raster or in vector format. The challenge of this problem is in determining the best process of mapping the surface information onto a segmented (street) network in order to guarantee a specific quality of the routing result. The process is carried out under the assumption that the surface data may change at regular intervals and that the intended routing process optimizes the path through the network by minimizing the cumulative cost of the traversed surface.

This approach differs significantly from determining a least cost path across a surface, in which the assumption is that the path can lead across any part of the surface. In our problem we deal with movement restricted to an existing network. It also differs from the classical A* optimization, because the underlying street segments need to be further segmented depending on artificial boundaries imposed on an otherwise continuous surface in order to produce weights for the segments.

In this study we will explore the different approaches and evaluate routing results with respect to qualitative differences. We expect this study to show the variability in the results induced through the diverse ways of computing and mapping the surface data onto the network data.

## 2. Methods

In this section we study three different approaches to mapping surface data onto a network. Each of these approaches is based on spatial overlay, but they differ in handling the surface data in the overlay computations and in the spatial representations. As case study, we use data of Particulate Matter (PM2.5) from 36 stations in Beijing, China (Yu Zheng et al. 2014) and street data from OpenStreetMap (OSM). The data from the stations is used to generate a raster representation with a cell size of 50 meters using Kriging as well as a vector representation using Voronoi cells.

## 2.1 Spatial overlay using Kriging surface

A general and simple approach for combining raster and line vector data is overlay. The intersection creates line data subdivided along the raster cells and includes the raster attributes of PM data (Figure 1):

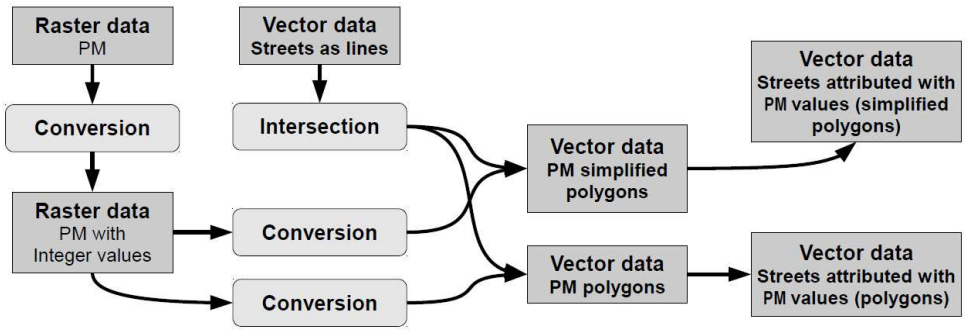



Figure 1. Conversion of raster into vector data and spatial join by intersecting vector datasets.

In case the raster data includes non-integer values, conversion of raster cells' data type into integer has to precede its conversion into polygon vector data. Both conversions lead to an approximation and joining of raster cells, achieving "*a polygon map of areas of similar characteristics*" (Congalton 1997: 426). Even though this procedure reduces the amount of raster data converted into polygons, accuracy of values is diminished, especially if polygons are simplified.

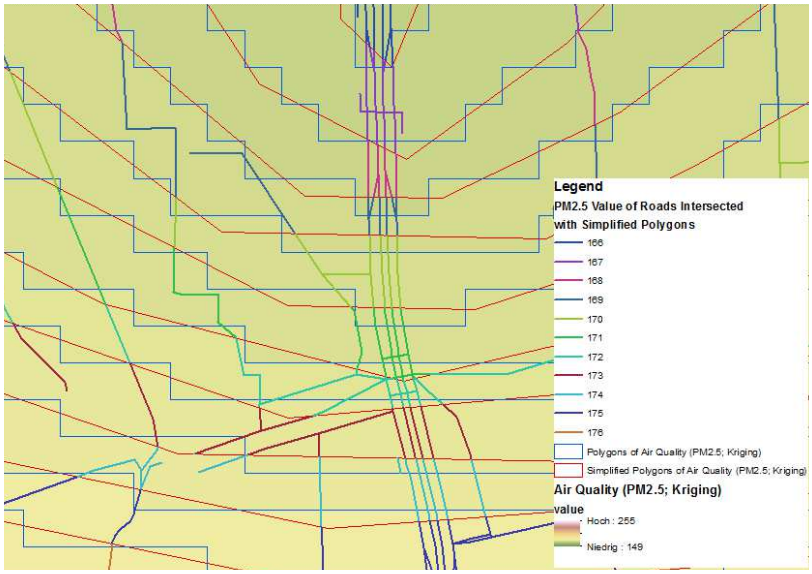



Figure 2. Comparison of Kriging interpolation results with/without simplified polygons.

Figure 2 demonstrates the differences of outcomes when intersecting streets with either polygons or simplified polygons. Differences of more than 25 meters can be detected, influencing the split of segments and finally the routing outcomes. Visualizing segments with PM2.5 attributes on the z-axis produces a step-like representation of the street data (see Fig.5).

### 3.2 Spatial Overlay using Voronoi Cells

Based on the original positions of the sensors for the measurement of PM2.5 values Voronoi cells categorize the road network by zones of influence. The position of the sensor (point) represents the centre of each Voronoi cell. The idea here is to provide fast results and not a detailed variety of interpolated values. Since the original PM2.5 values are used for this approach the values are equal for a large amount of road segments.

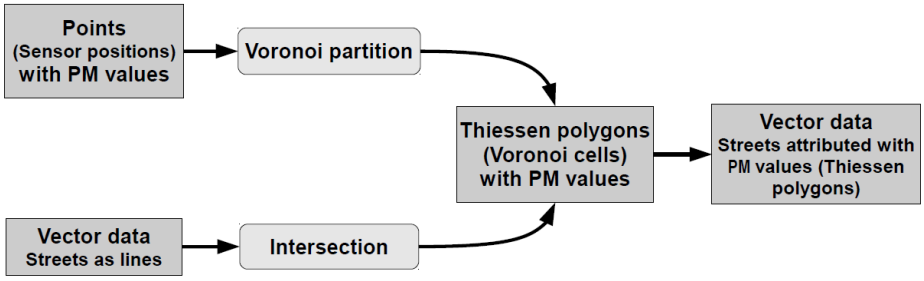



Figure 3. Creation of Voronoi polygons and the intersection with street segments.

Figure 3 shows the connection between the original street elements and the created Voronoi polygons. It is provided by the intersection of road segments and polygons. As can be expected the size of the Voronoi polygons varies greatly and directly influences the PM2.5 values of the street segments.

### 3.3 Spatial overlay using shape interpolation

In contrast to the approach in 3.1 it might be necessary to establish a connectivity between segments via the nodes, i.e., the nodes need to inherit the attribute of the line segments' Z value. This method therefore adds the Z values to each start and endpoint of a segment before converting the line segments into topological paths of a network (Fig. 4).

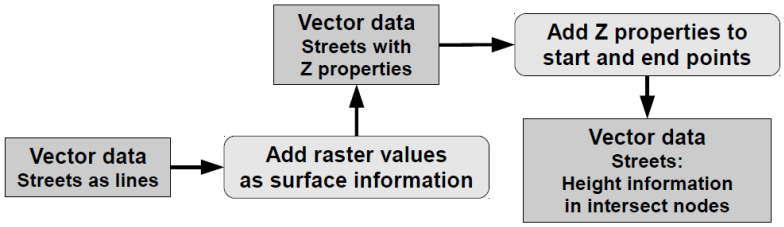



Figure 4. Interpolation of raster data as Z values of start and endpoints of vector data.

To achieve a higher detailedness of line segments, line data can be densified and split into smaller segments. Adding raster data as surface information then provides a higher degree of approximation to continuous data derived from raster data. However, the amount of line data grows depending on street length and consequently computing time increases. Merging line segments in case of similar Z values reduces the number of segments without degrading detailedness. Computing Z values for start and end point of each line segment not only establishes connectivity, but also contributes to further analysis such as slope and directions.

# 4. Results

The challenge of this problem is in determining the best process of mapping information given as a surface onto a network in order to guarantee a specific quality of the results. We explored several approaches to arrive at a network representation containing the information of interest in preparation for routing.

## 4.1 Segment height

Fig.5 shows the differing results with respect to where the height information of the original surface is stored. Storing the attribute values of the surface in the nodes of the new segments results in a detailed surface. Storing them in the original network nodes delivers a generalized "network surface". The Voronoi approach results in different segments compared to A in Fig.5 as well as in different values for the street segments, since only the original sample values are used.

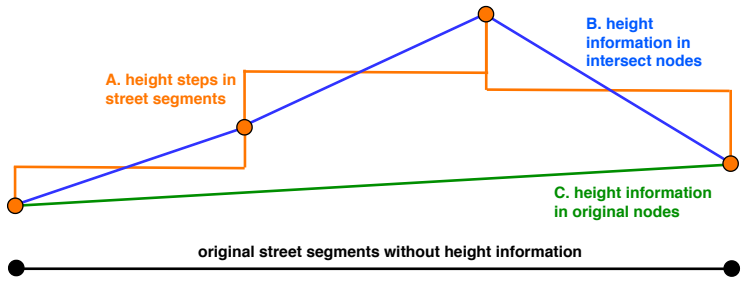



Figure 5. Differing results on segment height; ref. approach 3.1 (A.), 3.3 (B.); original street segments are retained (C.)

## 4.2 Routing

We are interested in evaluating the impact of the different ways of obtaining attribute values for the street segments for a routing problem. Routing through the networks resulting from the methods discussed above produces significantly different results (see Fig.6).

Table 1: Comparison of the routes' length and PM with regard to different used methods

| **Method** | **Overlay - raster polygons** (1a) | | **Overlay - simplified polygons** (1b) | | **Voronoi cells** (2) | | **Shape interpolation** (3a) | | **Densify and Shape interpolation** (3b) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | length [m] | PM | length [m] | PM | length [m] | PM | length [m] | PM | length [m] | PM |
| **Route A→B** | 115577.0 | 25763 | 116607.8 | 25710 | 140568.1 | 5495 | 142434.1 | 4636 | 72779.6 | 316264 |
| **Route C→D** | 20848.6 | 9395 | 20847.5 | 9356 | 19358.5 | 5469 | 22588.3 | 4839 | 18087.8 | 79180 |
| **Route E→F** | 33740.4 | 17615 | 33817.7 | 17417 | 31140.9 | 5751 | 31141.6 | 4804 | 26296.4 | 120707 |

Table 1 shows the differing measure of length and PM2.5 cumulative value of the path. It is directly discernible that the values differ significantly across the different methods. We lack the space here to discuss the differences in detail, a discussion will be engendered at the conference.

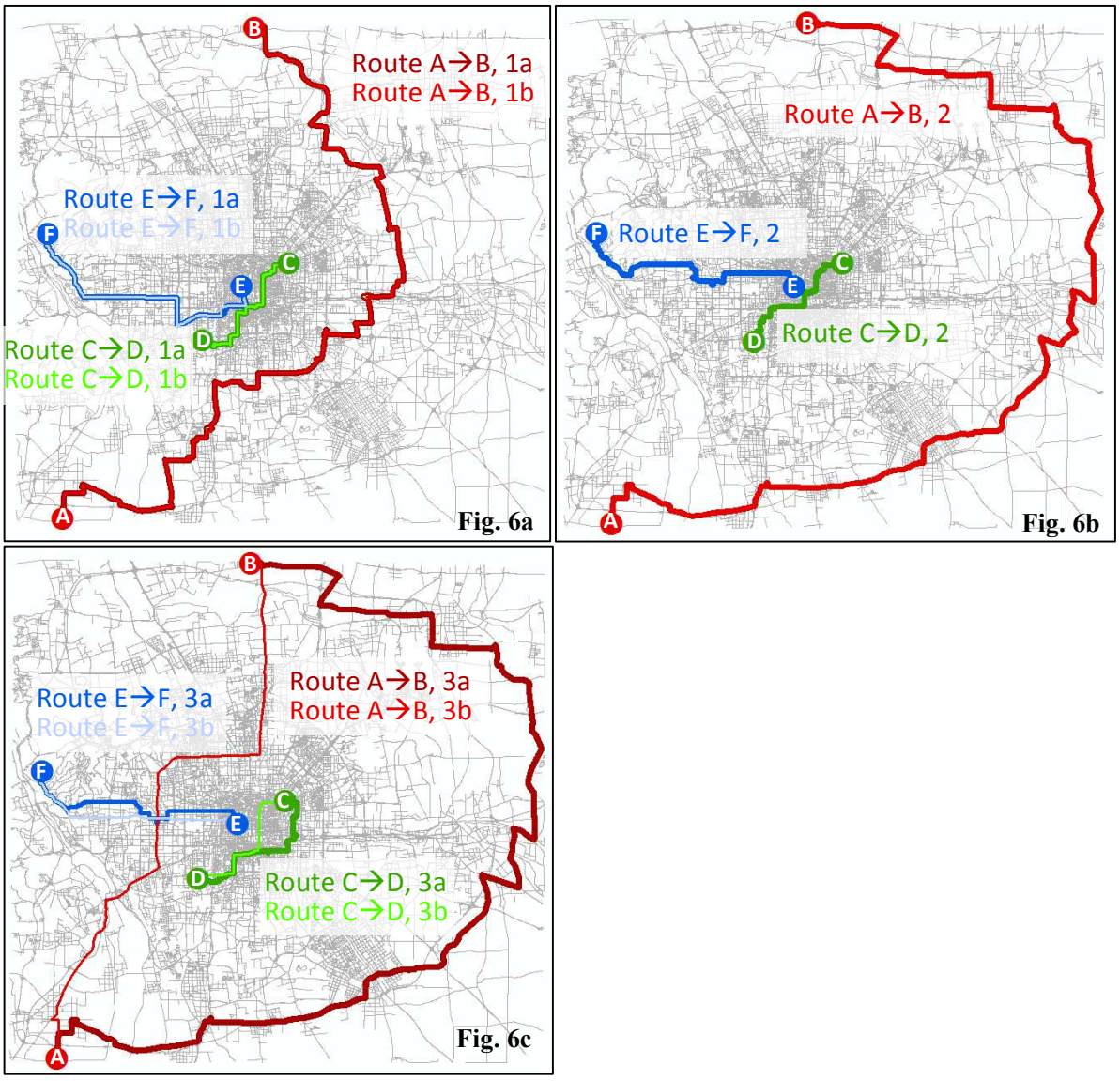



**Figure 6.** Comparison of routes

## References

Antikainen H, 2013, Using the Hierarchical Path Finding A* Algorithm in GIS to Find Paths through Rasters with Nonuniform Traversal Cost. ISPRS International Journal of Geo-Information, 2: 996-1014.

Congalton RG, 1997, Exploring and Evaluating the Consequences of Vector-to-Raster and Raster-to-Vector Conversion. Photogrammetric Engineering & Remote Sensing, Vol. 63, No. 4: 425–434.

Gehring U, Gruzieva O, Agius RM, Beelen R, Custovic A, Cyrys J, Eeftens M, Flexeder C, Fuertes E, Heinrich J, Hoffmann B, de Jongste JC, Kerkhof M, Klümper C, Korek M, Mölter A, Schultz ES, Simpson A, Sugiri D, Svartengren M, von Berg A, Wijga AH, Pershagen G and Brunekreef B, 2013, Traffic-related air pollution and lung function in children – the ESCAPE project. Environmental Health Perspectives 121(11-12): 1357-1364.

O'Sullivan D and Unwin DJ, 2010, Geographic information analysis. John Wiley & Sons, Inc., Hoboken, New Jersey.

Zheng Y, Chen X, Jin Q, Chen Y, Qu X, Liu X, Chang E, Ma W-Y, Rui Y and Sun W, 2014, A Cloud-Based Knowledge Discovery System for Monitoring Fine-Grained Air Quality. no. MSR-TR-2014-40, March 2014.

# Towards a better understanding of dynamic interaction metrics for wildlife: a comparison of null models

Jennifer A. Miller

The University of Texas at Austin. 1 University Station A3100 Austin, TX 78712 USA
Email: jennifer.miller@austin.utexas.edu

## 1. Introduction

Technological advancements in GPS and related satellite tracking technologies have resulted in significant increases in the availability of highly accurate data on moving objects, dramatically outpacing the development of appropriate methods with which to analyse them. Within GIScience, 'movement pattern analysis' (MPA) has emerged as a subfield that addresses concepts and theories used to explore the spatio-temporal structure in data in order to perform meaningful analysis, however the methodological and analytical framework associated with MPA is new and still evolving.

Interactions, for which the basic unit of observation is a pair of locations for two individuals, can be considered a second order property of movement but their social and psychological explanations and implications are far less generalizable. The nature of interactions between individuals of an animal population is a fundamental aspect of a species' behavioural ecology and information on the frequency and duration of these interactions is vital to understanding mating and territorial behaviour, resource use, and infectious disease epidemiology.

Interaction metrics used in wildlife studies have been classified as 'static' or 'dynamic' (Doncaster 1990). Static interaction metrics typically involve comparing the spatial overlap of the home ranges for two individuals to the rest of their respective home ranges and the interactions they describe are purely spatial or 'static' in that they do not account for the possibility of temporal avoidance or attraction between individuals. 'Dynamic interaction' between two individuals is defined as occurring within a spatial and temporal threshold and can provide information on possible attraction and avoidance of individuals that are in the same area at the same time (Doncaster 1990). Dynamic interaction rates are far more useful for understanding how two individuals interact in the context of disease transmission and behavioural ecology, but they are more problematic to measure.

In spite of the importance of interactions, they have not been a main research focus in either movement pattern analysis (GIScience) or movement ecology. Instead, research topics of interest to both sub-fields tend to 'follow' or accompany technological advancements. More recent studies using movement data to study interactions have also focused primarily on technological advancements related to the ability to measure interactions (e.g., 'proximity loggers', see Drewe et al. 2012) rather than the ability to interpret or understand them as meaningful indicators of level or degree of interaction. Additionally, few studies have tested different dynamic interaction metrics using the same data, and when they have been compared, the results have been quite variable (Miller, 2012). Most dynamic interaction metrics use an index or coefficient (e.g., ranging from 0 to 1, or -1 to +1) to describe both the magnitude and type (attraction, avoidance) of interaction, ostensibly incorporating a concept of neutral interaction or independent movement between individuals as a null model to which

observed values are compared. However, without realistic 'expected' values, none of these interaction metrics facilitates a more meaningful understanding of the nature of these interactions- e.g., whether the encounters occur more or less frequently than they would occur if the individuals were moving randomly across their range. Some dynamic interaction metrics are estimated using an expected value- although there is variation in how these expected values are calculated and what they might really represent.

## 2. Quantifying animal interactions

Many of these interaction metrics were originally developed in wildlife and behavioural ecology for use with direct observations of individuals (point-based) where observations were classified as one of four types: individual α and β together; α without β; β without α; or neither α nor β (see Cairns and Schwager, 1987 for review). Most of the interaction metrics described below have extended this concept by defining "together" either in terms of home range overlap or a spatio-temporal threshold.

Interactions have been studied in GIScience as a type of relative motion called 'reaction movement patterns' and defined as the "spatio-temporal footprint of the movement behaviour occurring when individuals react on their spatial co-occurrence" (Merki and Laube, 2012: 2). As opposed to many of the ecology-related interaction metrics based that use a spatio-temporal threshold, GIScience interaction metrics are more often concerned with measuring the similarity of movement parameters such as step length, direction, velocity, or more advanced descriptors such as sinuosity and are therefore considered path-based. Research in this context has focused on developing algorithms that detect pre-defined movement patterns such as "pursuit and escape", "confrontation" and "avoidance" and they have been applied to datasets of animal as well as human movement (e.g., soccer players).

While a few path-based dynamic interaction metrics have been introduced recently (see Long et al. 2014), path-based methods do not address relative spatial location and are more appropriate for measuring path similarity than the degree to which individuals encounter each other. As this work is primarily focused on measuring spatially proximal interaction, I focus on point-based metrics. Table 1 briefly describes the dynamic interaction metrics that are tested here, along with an indication as to whether they explicitly incorporate an expected value in their calculation.

The dynamic interaction metrics listed above are important ways to measure level of interaction between two individuals. However, each of them has sufficient limitations that prevent robust and meaningful analysis of interactions. Point-based interaction metrics often require the calculation or selection of highly subjective factors such as home ranges or a distance threshold ($d_c$). Most importantly, these metrics lack a benchmarking framework that deals with null models or expected values for neutral interaction in order to facilitate more meaningful interpretation of their values. The research presented here borrows from the null model approach commonly used in community ecology to compare observed (empirical) dynamic interaction values with distributions of expected values generated by using different null models.

Table 1: Point-based dynamic interaction metrics tested here.

| Metric | Description | interpretation | Reference |
|---|---|---|---|
| **Explicitly incorporates null expectation** | | | |
| Coefficient of sociality (**$S_c$**) | $\frac{D_E - D_O}{D_E + D_O}$ | -1 to 1 | Poole (1995) |
| Difference coefficient (c) | Extension of Doncaster test for distance intervals; $\frac{FO - FE}{FE}$ | Negative values suggest negative interaction; Positive values suggest positive interaction | White and Harris (1994) |
| Doncaster's nonparametric test (**Don**) | Frequency of observed s-t-matches (FO) compared to unmatched (expected; FE) | $Chi^2$ test | Doncaster (1990) |
| Minta's coefficient (**LIXN**) | $ln\left(\frac{\frac{o_{11}}{e_{11}} + \frac{o_{22}}{e_{22}}}{\frac{o_{12}}{e_{12}} + \frac{o_{21}}{e_{212}}}\right)$ | Positive values indicate joint use > solitary use; negative values indicate joint use < solitary use | Minta (1992) |
| **Does not explicitly incorporate null expectation** | | | |
| Half-weight association index (**HAI**) | $\frac{m}{m + 1/2(a + b)}$ | 0 to 1 | Cairns and Schwager(1987) |
| | | | |

## 3. Testing null models for interpreting dynamic interactions

Using GPS collar data from brown hyena dyads in Northern Botswana (see Miller 2012) this research explores the use of four different types of null models with which to compare the existing dynamic interaction metrics:

- **Shuffled coordinates-** refers to methods that involve using coordinate values of actual locations, but randomly shuffling them or measuring interactions for pairs of coordinates that did not actually occur at the same time.;
- **Random movement-** involves using different models of random movement based on an individual's movement parameters (see Miller 2012);
- **Rotated trajectories-** involves randomly rotating and shifting actual movement trajectories so that a path is maintained but it is located randomly in the study area.

Preliminary results indicate that these widely used dynamic interaction metrics are quite incongruous in terms of the type and degree of interactions that they measure. This is problematic as the ability to understand how individuals interact has important implications for understanding the spread of disease as well as behavioural ecology for less observable individuals.

## Acknowledgements

Thanks to Glyn Maude (Makgadikgadi Pans Brown Hyena Project), Paul Holloway (UT), and to the University of Texas at Austin Vice President of Research and the National Science Foundation (#0962198) for supporting this work.

## References

Benhamou S, Valeix M, et al. (2014) Movement-based analysis of interactions in African lions. Animal Behaviour, 90, 171–180.

Cairns SJ and Schwager SJ (1987) A comparison of association indices. Animal Behaviour, 35(5), 1454–1469.

Doncaster CP (1990) Non-parametric estimates of interaction from radio-tracking data. Journal of theoretical biology, 143(4), 431–443.

Drewe JA, Weber N, Carter SP, et al. (2012) Performance of Proximity Loggers in Recording Intra- and Inter-Species Interactions. PLoS ONE, 7(6), e39068.

Long JA, Nelson TA, Webb SL, et al. (2014, in press) A critical examination of indices of dynamic interaction for wildlife telemetry studies. Journal of Animal Ecology.

Merki, M and Laube, P (2012) Detecting reaction movement patterns in trajectory data. In: Proceedings of the *AGILE'2012 International Conference on Geographic Information Science*.

Miller JA (2012) Using Spatially Explicit Simulated Data to Analyze Animal Interactions: A Case Study with Brown Hyenas in Northern Botswana. Transactions in GIS, 16(3), 271–291.

Poole KG (1995) Spatial organization of a lynx population. Canadian Journal of Zoology, 73(4), 632–641.

White PCL and Harris S (1994) Encounters between Red Foxes (Vulpes vulpes): Implications for Territory Maintenance, Social Cohesion and Dispersal. Journal of Animal Ecology, 63(2), 315–327.

# Edge-based communities for identification of functional regions in a taxi flow network

U. Demšar[1], J. Reades[2], E. Manley[3], M. Batty[3]

[1]Centre for GeoInformatics, University of St Andrews, UK
Email: urska.demsar@st-andrews.ac.uk

[2]Geography Department, King's College London, UK
Email: jonathan.reades@kcl.ac.uk

[3]Centre for Advanced Spatial Analysis, University College London, UK
Email: {ed.manley, m.batty}@ucl.ac.uk

## 1. Introduction

Recent technological advances in spatial data collection have caused an explosion of new data volumes and their availability. One of these data types are flow networks, sometimes also called origin-destination (OD) networks which are now being increasingly captured using various forms of sensor technology from bespoke system which track vehicles and passengers to smart phone locations can that can be associated with individual travellers. These networks consist of vertices representing locations where flows start and end. Edges of the network bear information on the flow size and direction, thus forming a directed weighted network on spatial vertices. Examples of flow networks are transportation networks (e.g. flows of passengers between subway stations), migration/commuting networks and mobile phone communication networks.

In the geographic tradition which can be traced back forty years at least, one of the uses for flow networks has been in context of regionalisation: flow information was used to derive regions of functional interaction between origin and destination locations. Studies used a number of different flow data for this purpose: transportation flows (Black 1973), phone calls (Clark 1973, Goddard 1973), and taxi journeys (Goddard 1970). However, these studies were limited due to the limits on computer power and data available and perhaps due to this, network-based regionalisation seems to have been temporarily forgotten. Only recently has this topic received renewed attention: for example, regionalisation studies use commuting networks (Farmer and Fotheringham 2011, Landré and Håkansson 2013) and mobile phone communication networks (Expert et al. 2011, Thomas et al. 2012). This renewal is largely based on an interdisciplinary transfer of methods from network science research in physics, in particular various community detection methods that have been used to partition and summarise clusters that comprise such networks (Newman 2006).

In network science, a community is defined as a set of vertices in the network which are more densely interconnected with each other than with the rest of the network (Newman 2004). In geographic regionalisation, this corresponds to the concept of functional regions, which are spatially contiguous, internally well connected and relatively cohesive in terms of flows (Farmer and Fotheringham 2011). A number of algorithms have been developed for community detection in physics, where most approaches partition the vertex set to obtain communities. This means that each vertex can belong to at most one community and it is not possible for communities to share vertices with each other, which may be problematic in real-world networks (Palla et al. 2005).

In spatial flow networks the non-intersection criterion is in contradiction with the idea of polycentricity in movement, which is the concept that there exist several central locations which generate and receive large numbers of flows across a wide area and where trips from

each of the centres are not exclusively delimited from trips from all other centres. The polycentric movement process has been observed both at the level of mega cities as well as smaller areas (Hall and Pain 2006; Zhong et al., 2014). Indeed, the observation of the necessity of overlapping regions in movement-based regionalisation can again be traced back forty years to work by Goddard (1970) and his work on partitioning the centre of London into travel regions based on taxi flows. However, none of the recent network-based regionalisation approaches takes this overlapping necessity into consideration.

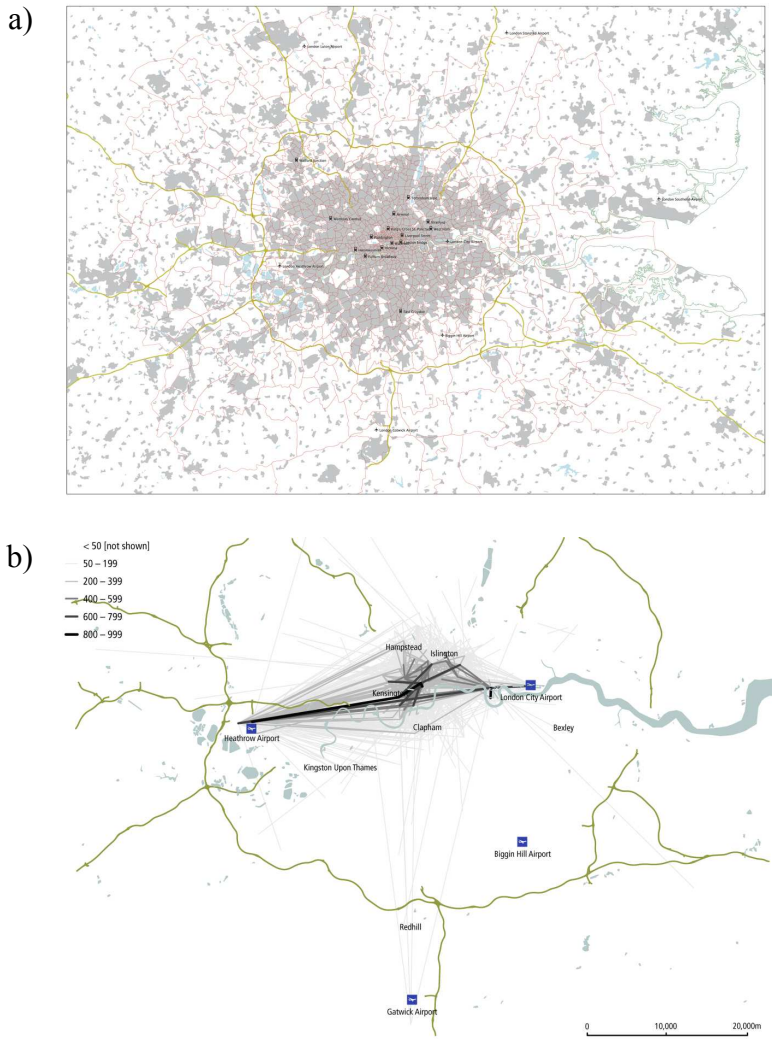



Figure 1: a) Traffic Area Zones (TAZes) in our study area. b) Taxi flows among TAZes.

In this paper we take inspiration from Goddard (1970) and investigate the possibility of using an edge-based community detection algorithm (Ahn et al. 2010) for identification of overlapping functional regions defining taxi flows in Greater London Area. This is work in progress: we demonstrate some initial results and discuss further directions for edge-based community detection in the context of spatial flow networks.

## 2. Data: taxi flows

For this study we were given access to three months of work day taxi flow data by Addison Lee minicabs (Dec 2010 – Feb 2011). Data consisted of GPS trajectories of taxis, and we aggregated origins and destinations of each trajectory into a set of Traffic Analysis Zones

(TAZ) to obtain a flow network. Figure 1 shows the TAZes covering the Greater London Area and the taxi flows.

We performed our analysis at two spatial scales: for Central and Inner London and for the Greater London Area. Table 1 presents the sizes of the two flow networks at these two spatial scales.

Table 1. Network sizes.

| | Number of vertices (TAZes) | Number of edges (non-zero flows between TAZes) |
|---|---|---|
| Central and Inner Area | 391 | 50,786 |
| Greater London Area | 1,165 | 104,587 |

## 3. Edge-based community detection

A typical community detection algorithm operating on vertices (Girvan & Newman 2002, Newman 2006) starts by calculating the similarity between each pair of vertices. Vertices are then aggregated using hierarchical clustering. This procedure starts with each vertex as representing one cluster. Vertices/clusters are then joined iteratively so that at each step the two clusters are joined that contribute the least to the increase in overall dissimilarity. This builds a dendrogram representing the temporal sequence as to how vertices/clusters are joined. A partition of the vertex set is obtained by cutting this dendrogram at some level. The best level is defined through optimisation of a modularity function, which reaches the maximum value when intra-cluster similarity is maximised and inter-cluster similarity is minimised. The resulting optimal partition splits the set of vertices into non-overlapping groups (communities), i.e. each vertex can only be a member of one of the groups.

Edge-based community detection (Ahn et al. 2010) operates in the same way as vertex-based community detection with one difference: it is the set of edges that is being partitioned rather than the set of vertices. Partitioning edges rather than vertices makes sense in cases where the network under consideration is a social network, since each vertex (a person) can belong to several not necessarily overlapping communities (social groups), e.g. colleagues, friends, family, etc. (Palla et al. 2005). This also makes sense for our taxi flow data, since it is reasonable to expect that each vertex (each TAZ) could feed flows of taxi traffic into several other TAZes, which do not necessarily feed flows among each other.

We start with a directed network of taxi flows between TAZs. Note that the Ahn et al. (2010) algorithm is suitable for undirected networks, whereas the flow networks are directed and weighted. We consider possible adjustments to directed networks as future work, while here we transform our flow network into an undirected one. This can be done in several ways, but we use the simplest and the most frequently used approach which is to sum the bidirectional flows and then discard direction (Leicht and Newman, 2008), which produces an undirected weighted network.

In the next step we calculate the similarity of each pair of connected edges (edges that share a common vertex) using a function that compares the topological structure of the neighbourhoods of both edges (i.e. edges that share a common vertex with the two original edges), the Tanimoto coefficient. This coefficient also uses edge weights (i.e. flow sizes) in the similarity calculation (see Ahn et al. 2010 for details)..

Once pair-wise similarity is calculated for all pairs of connected edges, we produce a dendrogram based on edge-similarity. Further we calculate the modularity function - partition density. Density of one community is a topological measure, defined by the number of links in the community, normalised with the maximum possible number of edges between the nodes in the community. Partition density is the average community density over all communities in one particular partition. This value has one global maximum between the top and the bottom of the dendrogram (Ahn et al. 2010) – this is at level k where the inter-

community density is maximal and intra-community density is minimal, thus defining the best possible partition of the edges into edge-communities.

In the final step we cut the dendrogram at level k, to obtain the best possible partitioning of edges into edge-communities.

a) b)

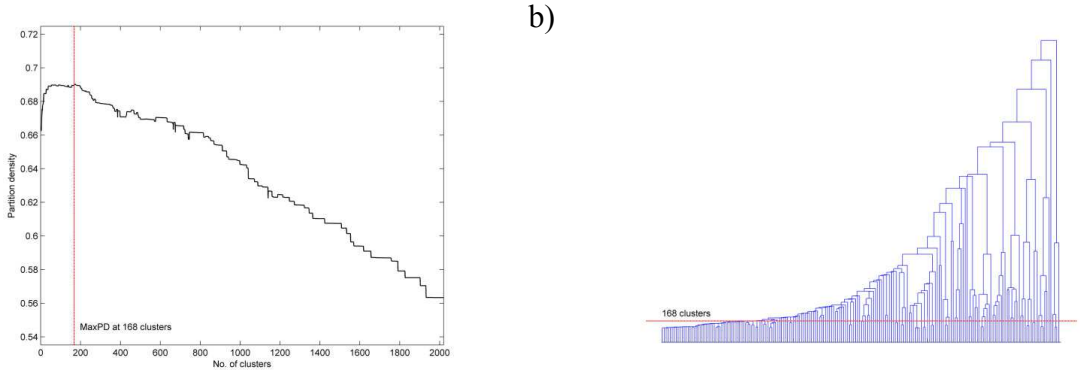



Figure 2: a) Partition density and b) dendrogram for the Central and Inner Area, cut at the optimal partition density level.

## 4. Results

The edge-based community detection algorithm splits our two taxi flow networks as follows: the optimal number of clusters is 168 for the Central and Inner Area, and 6,276 clusters for entire Greater London Area. Such large numbers of clusters in the optimal partition pose a particular challenge for interpretation of results, which we comment on in the discussion.

For illustration, figure 2 shows the maximisation of partition density and the resulting dendrogram cut for the 168 clusters in the Central and Inner Area. Further sorting the 168 clusters according to their size (number of edges included in each cluster), and taking into consideration only clusters containing more than 10 edges leaves 25 clusters as potential functional regions in taxi traffic in Central and Inner Area. These edge clusters and our tentative interpretation of the type of traffic they represent are shown in figure 3.

Addison Lee minicabs have a strong business bias: in contrast with London's more familiar black cabs they cannot be hailed on the street and need to be pre-booked. Further, the company prioritises customers with accounts, the majority of which are large businesses with extensive mobility requirements. As a consequence, we expect to see most of the flows to be for either business travel purposes or for trips to events. The resulting edge-communities (figure 3) appear at first exploration to be consistent with the expected business bias in the taxi traffic, but this bears further detailed investigation of results..

One of the surprises in our edge-based regionalisations is that the first cluster is vastly larger than all other clusters combined and that flows in this cluster do not seem to have any particular spatial distribution (note that clustering is purely data-driven and done on flow information only and therefore location does not play any role in cluster assignment). We speculate that this cluster may represents the `usual state of affairs' or, to put it in another way, the everyday traffic that encapsulates a dominant configuration that is both fairly uniform and geographically extended, but its size masks interesting secondary functional groupings in taxi flows, which we can identify from the rest of the communities.

The results however bear further investigation and perhaps algorithm validation on a synthetically generated flow network, where we would be able to control for patterns that we would expect to be able to identify with our method.

We further encounter a "big data" problem when attempting to partition the whole Greater London Area. This partition is optimal when 104,587 edges are split into 6,278

clusters (fig. 4) and we are contemplating possibilities as to how to visualise let alone interpret such a large partition.

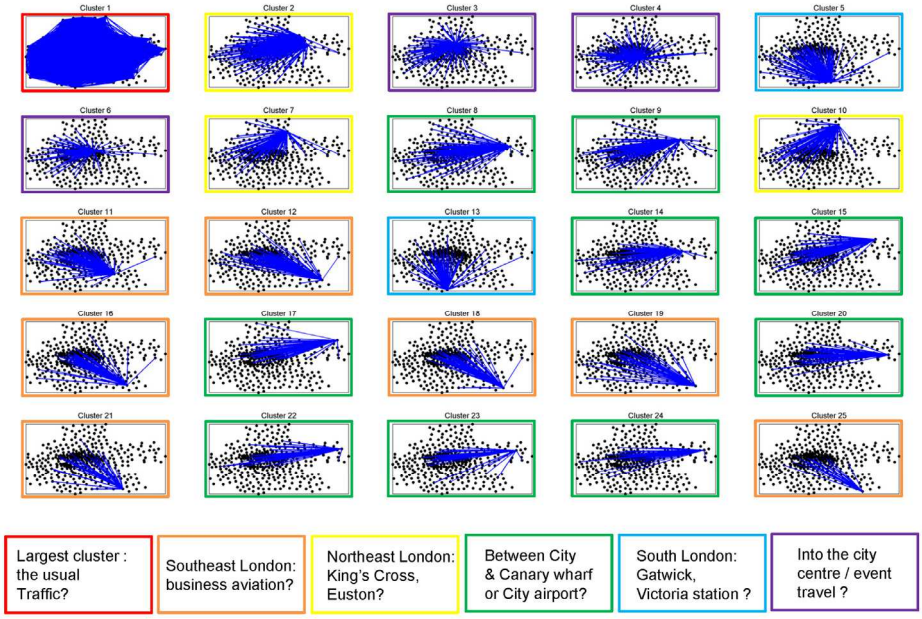



Figure 3: Tentative categorisation of the largest edge-based taxi flow clusters for Central and Inner Area. Interpretations shown are only possibilities and need to be further investigated.

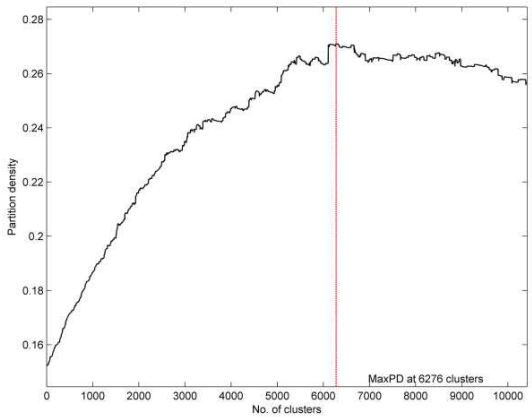



Figure 4: Partition density max for the Greater London Area. is reached at 6278 clusters.

## 5. Conclusions and outlook

This paper presents an attempt at edge-based regionalisation of taxi traffic in London and is a work in progress. As mentioned above, one of the major problems we encounter is that while we demonstrate that the algorithms taken from complex networks research can be potentially applied on real-life spatial flow networks with a tentatively plausible geographical results, the size of results may be prohibitory towards proper understanding and interpretation of results (e.g. 6278 clusters in taxi traffic in Greater London Area). The task of interpreting thousands of clusters resulting from the mathematically-derived optimal partition is demanding and further consideration will be needed to address it. The problem is not limited to our case: most of flow data sets are very large and can produce a large number of optimised regions in the best partition. For example, UK migration flow data are collected between 223060 census

output areas, which generates a network with more than 49 million edges to be classified into regions. GIScience can not solve problems of this type on its own – we believe that interdisciplinary knowledge exchange of spatial sciences with other disciplines that deal with network-style big data (such as physics and computer science) would be necessary to approach these problems.

Further, in this attempt we implemented an algorithm originally developed for non-spatial undirected networks. How to include direction in this process is a topic for future research. In flow networks, space also matters and this needs to be recognised in the regionalisation procedure. Space can be taken into account in different ways, either by incorporating spatial autocorrelation into community-detection (Cerina et al. 2012) or by using a geo-aware modularity function (Hanningan et al. 2013). The second point is particularly relevant, since currently the best partition obtained through optimisation of partition density is not linked to any spatial properties of the flows, but only to topology of the network. There is an implicit inclusion of the sizes of the flows in the procedure through Tanimoto coefficient, which uses flow sizes as weights in the calculation of edge similarity, but locations of edges/vertices and other spatial properties are currently not considered. We plan to investigate these possibilities further and explore how these or similar novel space- and/or direction-aware methods can be used for improved regionalisation from flow networks.

## Acknowledgements

The authors would like to acknowledge support from Addison Lee and Transport for London (TfL) who contributed data for this study.

## References

Ahn YY, Bagrow JP and Lehmann S, 2010, Link communities reveal multi-scale complexity in networks. *Nature*, 466:761-765.

Black WR, 1973, Toward a Factorial Ecology of Flows. *Economic Geography*, 49(1):59-67.

Cerina F, De Leo V, Barthelemy M and Chessa A, 2012, Spatial Correlations in Attribute Communities. *PLoS One*, 7(5): e37507.

Clark D, 1973, Normality, transformation and the principal components solution. *Area*, 5:110-113.

Expert P, Evans TS, Blondel VD and Lambiotte R, 2011, Uncovering space-independent communities in spatial networks. *PNAS*, 108(19):7663-7668.

Farmer CJQ and Fotheringham AS, 2011, Network-based functional regions. *Environment and Planning A*, 43(11):2723-2741.

Girvan M and Newman MEJ, 2002, Community structure in social and biological networks. *PNAS*, 99(12):7821-7826.

Goddard JB, 1970, Functional Regions within the City Centre: A Study by Factor Analysis of Taxi Flows in Central London. *Transactions of the Institute of British Geographers*, 49:161-182.

Goddard JB, 1973, Office linkages and location: A study of communications and spatial patterns in Central London. *Progress in Planning*, 1:109-232.

Hall P and Pain K, 2006, *The polycentric metropolis: learning from mega-city regions in Europe*. Earthscan.

Hannigan J, Hernandez G, Medina RM, Roos P and Shakarian P, 2013, Mining for Spatially-Near Communities in Geo-Located Social Networks. *Social Networks and Social Contagion, AAAI Technical Report* FS-13-05.

Landré M and Håkansson J, 2013, Rule versus Interaction Function: Evaluating Regional Aggregations of Commuting Flows in Sweden. *European journal of transport and infrastructure research* 13(1):1-19.

Leicht EA and Newman MEJ, 2008, Community structure in directed networks. *Physical Review Letters*, 100:118703-1-118703-4.

Newman MEJ, 2006, Modularity and community structure in networks. PNAS, 103(23):8577-8582.

Palla G, Derenyi I, Farkas I and Viscek T, 2005, Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814-818.

Thomas I, Cotteels C, Jones J and Peeters D, 2012, Revisiting the extension of the Brussels urban agglomeration: new methods, new data… new results? *Belgeo* 1-2:1-12.

Zhong C, Arisona SM, Huang X, Batty M and Schmitt G, 2014. Detecting the dynamics of urban structure through spatial network analysis, *International Journal of Geographic Information Science*, forthcoming.

# Using mobile phone data to map human population distribution

C. Linard[1], P. Deville[2], M. Gilbert[1], V.D. Blondel[2], A.J. Tatem[3]

[1] Biological Control and Spatial Ecology, Université Libre de Bruxelles, CP 160/12, Avenue FD Roosevelt 50, B-1050 Brussels, Belgium
Email: {catherine.linard; marius.gilbert}@ulb.ac.be

[2] Department of Applied Mathematics, Université catholique de Louvain, Avenue G. Lemaitre, 4, 1348 Louvain-la-Neuve, Belgium
Email: {pierre.deville; vincent.blondel}@uclouvain.be

[3] Department of Geography and Environment, University of Southampton, Southampton, UK
Email: andy.tatem@gmail.com

## 1. Introduction

Many applications rely on information about the spatial distribution of human population, yet, our knowledge of human population distribution remains surprisingly poor in many areas of the world. Whilst the use of GPS and GIS in census data collection and processing, and the advent of detailed satellite imagery are facilitating improvements in spatial resolution and accuracy of population maps (Linard et al. 2012; Azar et al. 2013; Stevens et al. 2014), they remain tied to the census date and little information exists to inform on temporal changes in population distributions across scales of days, weeks, months or years. Such features constrain the effective application of population maps in situations where timely information is required, such as disasters, conflicts or epidemics.

The proliferation of mobile phones (MPs) offers an unprecedented solution to this data gap. The global MP penetration rate reached 96% in 2013 (International Telecommunication Union 2013). In developed countries, the number of MP subscribers surpasses the total population, with a penetration rate now reaching 128%, while in developing countries it is as high as 89%, and continuing to rise (International Telecommunication Union 2013). MP networks, also called cellular networks, are composed of cells, i.e. geographic zones around a phone tower. Each MP communication can be located by identifying the geographic coordinates of its transmitting tower and the associated cell. This network-based positioning method is simple to implement and its accuracy directly depends upon the network structure, the higher the density of towers, the higher the precision of the MP communication geo-localization (Mateos and Fisher 2006). Detailed records of MP calls and text messages therefore provide a valuable resource on where and when people are sending or receiving information and the widespread and increasing use of MPs offers a promising alternative data source for better understanding patterns and processes in human geography (Ahas et al. 2008; Yuan, Raubal, and Liu 2012; Järv et al. 2012).

Our objective here was to develop an approach that makes use of MP data to map the spatio-temporal distribution of human population over large spatial extents, but yet at high resolution. Our underlying assumption is that the number of phone calls transmitted through a tower scales with the number of people in its coverage area. In order to be widely applicable, the methodologies were designed to be easy to implement, while minimizing the impact of phone usage heterogeneities among social groups, regions and network providers. Using France as case

study, we show how aggregated MP data can be used efficiently to map population distributions and reveal otherwise unmeasurable patterns in space and time.

## 2. Materials and methods

A large dataset of MP calls obtained from a major carrier in France was used as a proxy for population activity in the country. The dataset covers a period of 5 months from May 2007 to October 2007, and contains more than 1 billion MP calls from 17 million users, which was approximately 30% of the population of metropolitan France in 2007. We used the aggregated number of MP calls made or received from/to the towers, without any individual information because this information was not available. This has two benefits: (i) it ensures that our population density estimation method requires only data that is readily collected and stored by network providers for billing purposes and (ii) the privacy of network customers is preserved.

The coverage area of towers was approximated using a Voronoi-like tessellation and the MP call density was computed for each Voronoi polygon (Figure 1B). The population density ($\rho_c$) in a given area $c$ was estimated as a function of the MP call density per day ($\sigma_c$) for that area using Equation 1:

$$\rho_c = \alpha \sigma_c^{\beta} \qquad (1)$$

where the parameters $\alpha$ and $\beta$ were fitted by a linear regression based on training data. The parameter $\alpha$ represents the scale ratio and $\beta$ the super linear effect of population density $\rho_c$ on the MP call density $\sigma_c$.

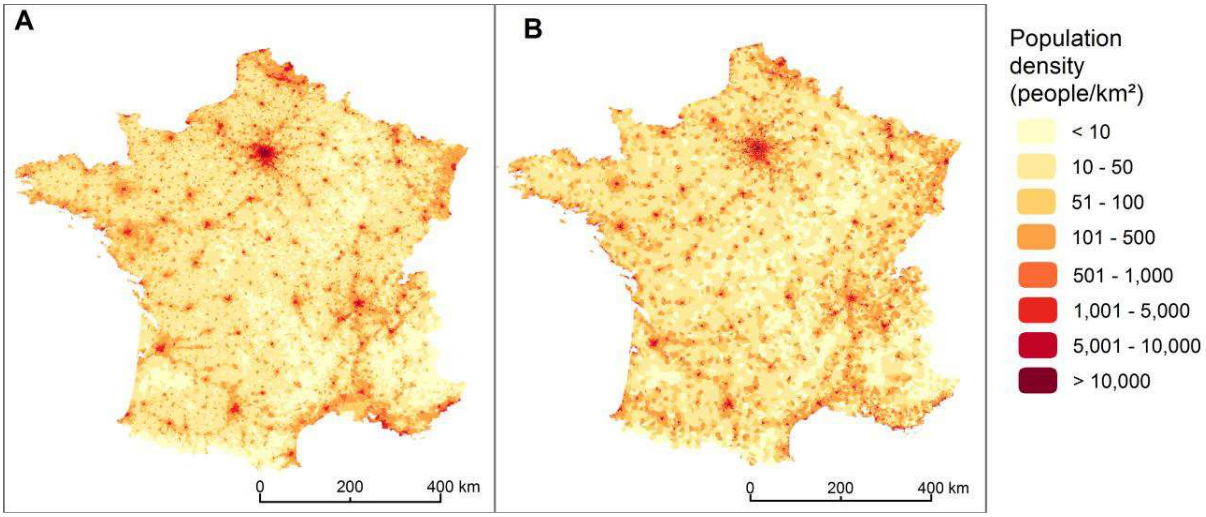



Figure 1: Comparison of (A) census-derived population density at the administrative unit level 5 (communes) with (B) MP-derived population density at the level of Voronoi polygons.

In order to assess the accuracy of the method, we compared the datasets produced with other widely used population distribution datasets such as the Gridded Population of the World (GPW) database (Tobler et al. 1995), the Global Rural Urban Mapping Project (GRUMP) (Balk and Yetman 2004), the LandScan Global Population database (Dobson et al. 2000; Bhaduri et al. 2007) and the WorldPop database (Stevens et al. 2014). While GPW and GRUMP mainly rely on a simple areal weighting scheme, Landscan and WorldPop incorporate a wide range of remotely-sensed and other geospatial data. The different population distribution datasets were

compared to baseline census data at the finest available administrative unit level (ADM-5) (Figure 1). However, a precise and quantitative accuracy assessment of the MP method would require daytime population data as reference, instead of census-derived nighttime data.

Temporal dynamics were derived from MP data using the timestamp associated to each MP call. Weekly dynamics were analyzed by dividing the MP data into calls performed during weekdays (Monday to Friday) and weekends (Saturday and Sunday), and seasonal dynamics were explored by dividing MP data into calls performed during the holiday period (July and August) and working periods (May, June, September and October). Predicted population densities for each unit and for both time periods were computed and relative differences between the two time periods were extracted.

## 3. Results and discussion

Globally, the population downscaling method based on MP data was found to be of comparable accuracy to existing downscaling methods. In urban areas, where the density of phone towers is high, the MP-based method captured a significant spatial variability in population density that was not captured by other methods.

The potential of MP data to estimate population density variations through time is illustrated in Figure 2. The relative differences in estimated population densities between the major holiday period and more traditional working periods reveal clear spatial patterns. Most cities are characterized by a large decrease of population densities during holidays, while less populated areas and well-known tourist sites show large increases.

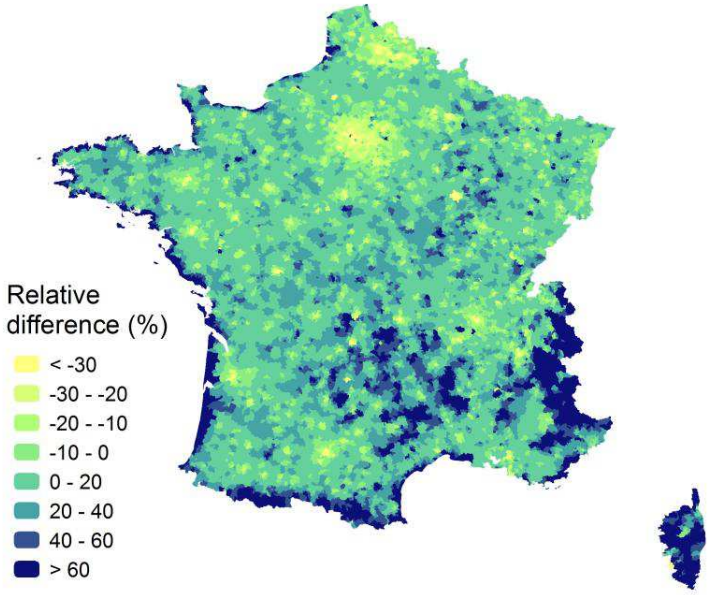



Figure 2: Relative difference in population densities predicted using mobile phone data by commune (ADM-5) between the main holiday period (July and August) and the working period (May, June, September and October).

Here, we use detailed records of MP calls to produce accurate and cost-effective datasets depicting human population distribution over large spatial extents. By using phone call activities at cell towers, we show how spatially and temporarily explicit estimations of population densities

across countries and their changes over multiple timescales can be produced, while preserving the anonymity of individual users. Methods require minimal input data and are therefore widely applicable. In addition, while socio-economic or demographic factors may bias population density estimates, preliminary analyses show that the impact of spatio-temporal variabilities in phone usage behaviours on population estimates is marginal. Comparisons with existing population mapping methods reliant on remotely sensed and other geospatial data revealed the high accuracy and flexibility of the phone-based approach. Provided that access to anonymized MP data becomes facilitated and more easily accessible to the scientific community, the prospect of being able to study human population distribution and movements over relatively short time intervals paves the way to new applications and near real-time understanding of large-scale patterns and processes in human geography.

## Acknowledgements

Authors thank Samuel Martin, Forrest R. Stevens and Andrea E. Gaughan for their contribution to this work. CL, PD and MG are supported by the Fonds National de la Recherche Scientifique (F.R.S./FNRS), Brussels, Belgium; part of this work was supported by the FNRS PDR T.0073.13. AJT is supported by funding from NIH/NIAID (U19AI089674), the Bill & Melinda Gates Foundation (49446, 1032350), and the RAPIDD program of the Science and Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. This work forms part of the WorldPop Project (www.worldpop.org.uk) and Flowminder (www.flowminder.org). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

Ahas, Rein, Anto Aasa, Antti Roose, Ülar Mark, and Siiri Silm. 2008. “Evaluating Passive Mobile Positioning Data for Tourism Surveys: An Estonian Case Study.” Tourism Management 29 (3): 469–86.

Azar, Derek, Ryan Engstrom, Jordan Graesser, and Joshua Comenetz. 2013. “Generation of Fine-Scale Population Layers Using Multi-Resolution Satellite Imagery and Geospatial Data.” Remote Sensing of Environment 130: 219–32. doi:10.1016/j.rse.2012.11.022.

Balk, Deborah L, and Greg Yetman. 2004. The Global Distribution of Population: Evaluating the Gains in Resolution Refinement. Center for International Earth Science Information Network (CIESIN), New York. Available at: http://sedac.ciesin.org/gpw/docs/gpw3_documentation_final.pdf.

Bhaduri, Budhendra, Edward Bright, Phillip Coleman, and Marie Urban. 2007. “LandScan USA: A High-Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics.” GeoJournal 69 (1): 103–17. doi:10.1007/s10708-007-9105-9.

Dobson, J. E, E. A Bright, P. R Coleman, R. C Durfee, and B. A Worley. 2000. “LandScan: A Global Population Database for Estimating Populations at Risk.” Photogrammetric Engineering and Remote Sensing 66 (7): 849–57.

International Telecommunication Union. “The World in 2013: ICT Facts and Figures.” http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx.

Järv, Olle, Rein Ahas, Erki Saluveer, Ben Derudder, and Frank Witlox. 2012. “Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records.” PLoS ONE 7 (11): e49171. doi:10.1371/journal.pone.0049171.

Linard, Catherine, Marius Gilbert, Robert W. Snow, Abdisalan M. Noor, and Andrew J. Tatem. 2012. “Population Distribution, Settlement Patterns and Accessibility across Africa in 2010.” Edited by Guy J-P. Schumann. PLoS ONE 7 (2): e31743. doi:10.1371/journal.pone.0031743.

Mateos, Pablo, and Peter F. Fisher. 2006. “Spatiotemporal Accuracy in Mobile Phone Location: Assessing the New Cellular Geography.” In Dynamic & Mobile GIS: Investigating Change in Space and Time, Jane Drummond, Roland Billen, Elsa João and David Forrest, 189–212. Taylor & Francis.

http://www.casa.ucl.ac.uk/pablo/papers/Mateos%20%26%20Fisher%20(2007)%20New%20Cellular%20Geography.pdf.

Stevens, Forrest R, Andrea E Gaughan, Catherine Linard, and Andrew J Tatem. 2014. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Other Ancillary Data." Plos One In press.

Tobler, Waldo, Uwe Deichmann, Jon Gottsegen, and Kelly Maloy. 1995. The Global Demography Project (Technical Report TR-95-6). Santa Barbara, CA: National Center for Geographic Information and Analysis, Department of Geography, University of California Santa Barbara.

Yuan, Yihong, Martin Raubal, and Yu Liu. 2012. "Correlating Mobile Phone Usage and Travel Behavior – A Case Study of Harbin, China." Computers, Environment and Urban Systems, Special Issue: Geoinformatics 2010, 36 (2): 118–30. doi:10.1016/j.compenvurbsys.2011.07.003.

# Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area

Song Gao[1], Jiue-An Yang[1,2], Bo Yan[1], Yingjie Hu[1], Krzysztof Janowicz[1], Grant McKenzie[1]

[1]STKO Lab, Department of Geography, University of California, Santa Barbara, CA, USA
Email: {sgao, jiueanyang, boyan, yingjiehu, jano, grant.mckenzie}@geog.ucsb.edu
[2]Department of Geography, San Diego State University, CA, USA

## 1. Introduction

Trajectory-based mobility research plays an increasing role in GIScience and related domains. Typically, the research results heavily depend on the quality and resolution of data that can be collected, e.g., via surveys. In travel behaviour and transportation studies, time and cost constrains are the limiting factors for the collection of large-scale individual travel behaviour data using traditional trip-diary surveys (McNally 2000). With the fast development of information and communication technologies (ICT), new data sources including GPS logs, smart card records, mobile phone data, and location-based social media have become potential alternatives or complementary approaches to study large-scale human mobility patterns and travel behaviour (Calabrese et al. 2011, Liu et al. 2012a, Yue et al. 2014). Human movement origin-destination (OD) information is of major importance in urban transportation modelling and infrastructure planning in order to optimize the use of street networks. The increasing use of social media like Twitter offers unprecedented opportunities to study individual activities, to know where users are at which time, and what they are talking about. In this work we study the reliability of detecting regional OD trips from individual geotagged tweets in comparison with survey data in a quantitative manner, and explore the spatiotemporal flow patterns extracted from social media.

We will investigate the research question of whether **OD trips mined from social media yield comparable results to expensive and labour intensive large-scale studies.** To do so, we will derive OD trips from geotagged tweets, aggregate them, and compare the results by correlating them to the American Community Survey data.

## 2. Data and Methods

### 2.1 Datasets

We collected 6.8 million geotagged tweets from 110,868 users in the Greater Los Angeles Area from December 7, 2013 to January 7, 2014. This area sprawls over five counties in the southern part of California, namely Los Angeles, Orange, San Bernardino, Riverside, and Ventura counties. We only use geotagged tweets whose sources are smart phones, including iPhone, Android, Blackberry and Windows Phones. This ensures that a geotagged-tweet reflects a person's physical location instead of a social-bot IP address or a default (hometown) location. Some initial data processing reveals that on average a user generates two geotagged tweets per day within the collection period. However, about 11000 (i.e., 10%) users tweet more than 5 geotagged tweets per day. Figure 1 shows that the distribution of the daily average number of geotagged tweets per user actually fits a truncated power function with the exponent value 1.94 and R-square 0.93. We also found the mean of individual average inter-tweeting time interval per day for all users to be 126 minutes, and the median is 79 minutes. In addition, as shown in Figure 2, the distribution of average inter-tweeting time interval per user varies from minutes to hours, and the majority (about 80%) of users is within 190 minutes per day. These preliminary analysis results help us understand the characteristics of geo-tweeting behaviours in our study area and guided us in the setting temporal-bands in the OD trip estimation algorithm discussed below.

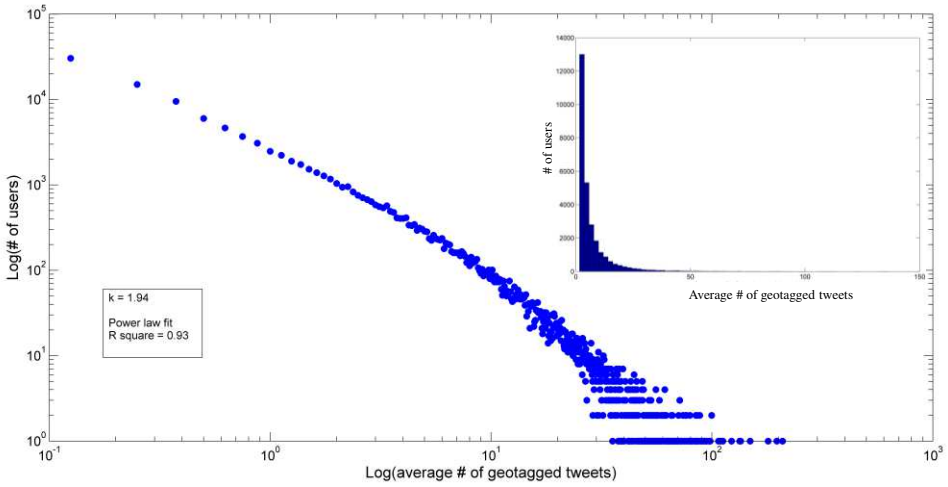



Figure 1: The log-log plot and histogram for the average number of geotagged tweets per user per day.

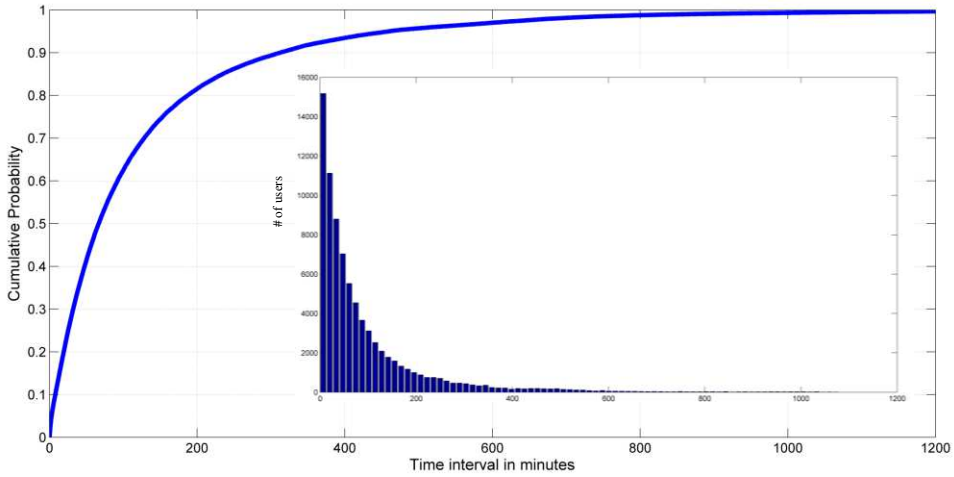



Figure 2: The histogram and cumulative probability distribution of individual average inter-tweeting time interval per day for all users

## 2.2 OD-Trip Estimation Approach

OD-trip estimation has been widely used for predicting travel demands in the conventional four-step model of transportation forecasting process. Our approach for estimating travel OD trips from geotagged tweets consists of two steps: individual-based trajectory detection and place-based trip aggregation.

In the first step, let $L_u = (l_{u,}^1 l_{u,}^2 ..., l_u^n)$ denote the temporal sequence of geotagged-tweet locations (latitude/longitude) of the user u. Then, we spatially joined all locations to the traffic analysis zones (TAZs) based on computing point-in-polygon relations which creates a second sequence $Z_u = (z_{u,}^1 z_{u,}^2 ..., z_u^n)$ of the user's location records at the TAZ scale. The spatial extent of a TAZ varies, ranging from large areas in the suburbs to as small as city blocks in central business districts. However, even for these small TAZs, the minimal extends of the bounding rectangles are about 600-1000 meters which is sufficient to filter the smartphone GPS uncertainty (typically up to 30 meters in our dataset). As s user might have multiple geotagged tweets within the same TAZ over a short period of time, these records do not contribute to the physical movements at the inter-TAZ level. Therefore, we spatially clustered those consecutive points if they were located inside the same TAZ polygon within the time threshold $\Delta t$ which we set to 4 hours based on the knowledge from aforementioned inter-tweeting time analysis. The new sequence of TAZ clusters can be represented as

$C_u = (c_u^1 c_u^2 ..., c_u^n)$. Next, individual trips can be extracted as the paths between two consecutive clusters in different TAZs for any given user.

In the second step, we aggregate trips (u, o, d, t) with the same origin o and destination d TAZ regions for all users together at different temporal windows t such as hourly, daily, or weekly. The result is an asymmetric OD matrix whose element $T_{ij}$ represents the total number of detected trips from the origin i to the destination j regions starting within a time period.

# 3. Results and Evaluation

## 3.1 Extracting Peak-Hour OD Trips at the TAZ-scale

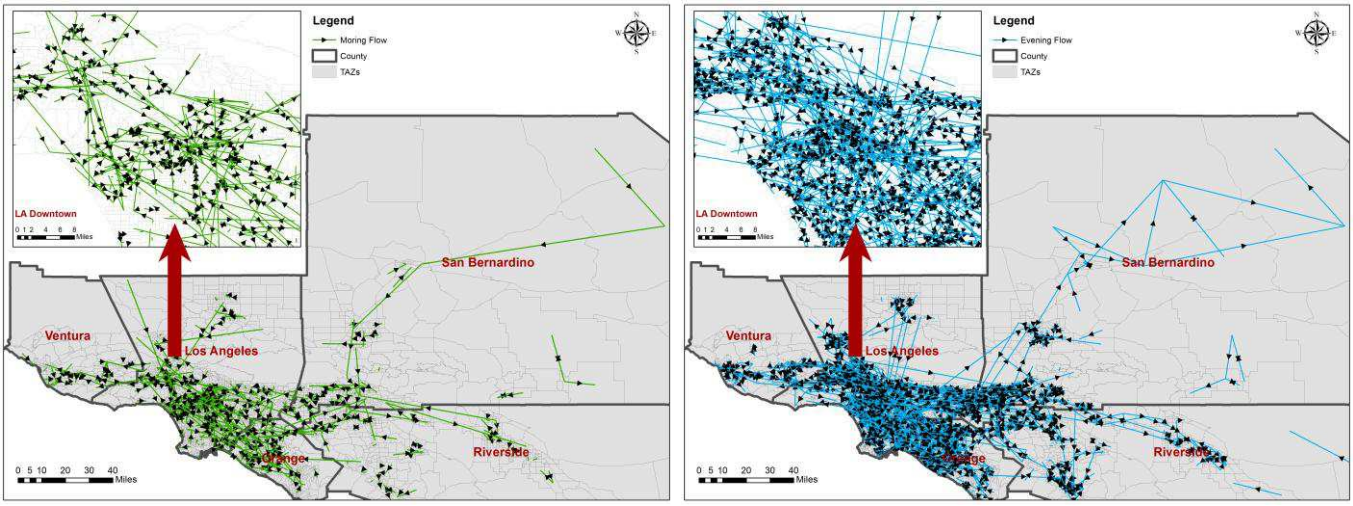
The proposed OD-trip estimation approach has the flexibility to detect dynamic inter-TAZ mobility flows at different temporal windows. In order to compare the detecting results with 2008-2012 American Community Survey (ACS)[1] data for evaluation, we aggregate OD trips extracted from geotagged tweets in 30min time windows based on the leaving time of each trip in morning-peak hours 5am-9am as shown in Table 1. On average, we detected about 24000 daily trips and the Pearson correlation coefficient between the survey data and the detected trips in weekdays is 0.91 (p-value 0.0017), a little lower for weekends 0.69 (p-value 0.05), and substantially lower for Christmas Day 0.59 (p-value 0.1233). The higher correlation between weekday trips and the survey at such a significance level than weekends and holidays meets our expectation since weekday trips have more regular patterns. Furthermore, we analyzed the trip-length distribution and found that it roughly follows a distance-decay distribution (Figure 3c and 3d) and the average length is about 56 km (35 miles). If we convert the trip distance into time using the local speed limit of 65 miles, the average time of all morning trips is about 32 minutes and very close to the survey data results of 29 minutes. All these results indicate that our OD-trip detection algorithm corresponds well with ACS data and can capture the overall characteristics of mobility flows in the study area using a big-data-driven approach.

In addition, the geovisualization of morning or evening peak-hour trips and netflow (inflow-outflow) patterns help us to identify the directed-flow changes in suburbs and downtown areas, as well as to better understand urban transportation dynamics (see Figure 3). Advanced spatiotemporal patterns and the linkages to land-use types can also be analyzed using the methods proposed by Guo et al. (2012) and Liu et al. (2012b) for further studies.

Table 1. The comparison of average morning peak-hour trips between the survey and the results detected from geotagged tweets

| Time Window | Survey | Weekdays | Weekends | Christmas |
|---|---|---|---|---|
| 5:00am – 5:29am | 6.74% | 2.31% | 3.92% | 5.68% |
| 5:30am – 5:59am | 7.12% | 4.09% | 5.22% | 9.61% |
| 6:00am – 6:29am | 13.18% | 8.53% | 7.65% | 9.61% |
| 6:30am – 6:59am | 12.36% | 15.29% | 10.63% | 11.35% |
| 7:00am – 7:29am | 20.40% | 24.80% | 16.98% | 12.66% |
| 7:30am – 7:59am | 14.92% | 20.89% | 23.69% | 14.41% |
| 8:00am – 8:29am | 16.99% | 15.73% | 17.72% | 23.58% |
| 8:30am – 8:59am | 8.28% | 8.36% | 14.19% | 13.10% |

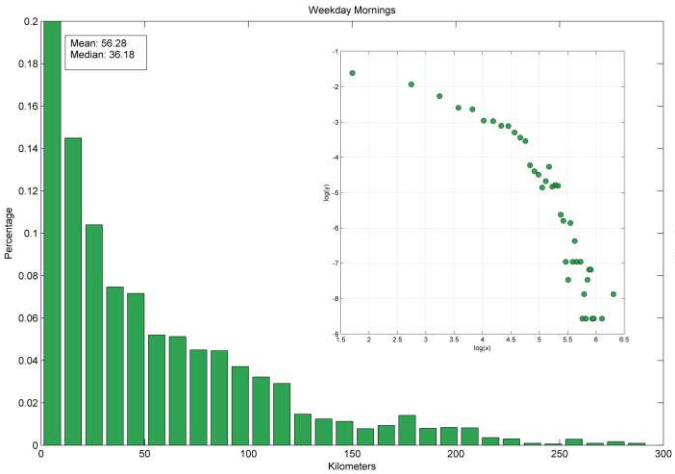
[1]Search for Table S0801: Commuting Characteristics by Sex via http://www.census.gov/acs/www/



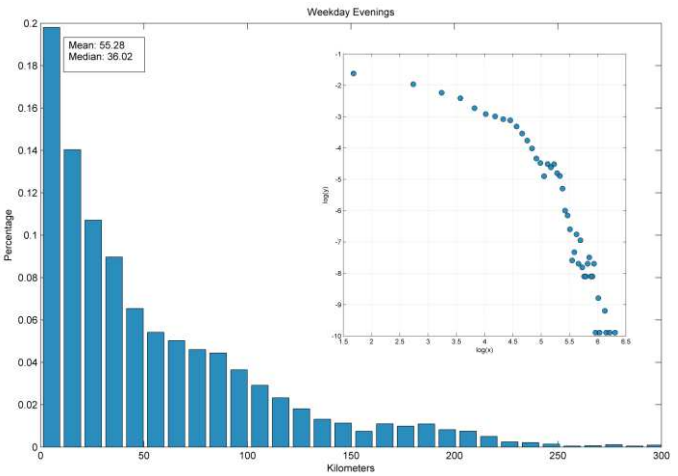

(a)



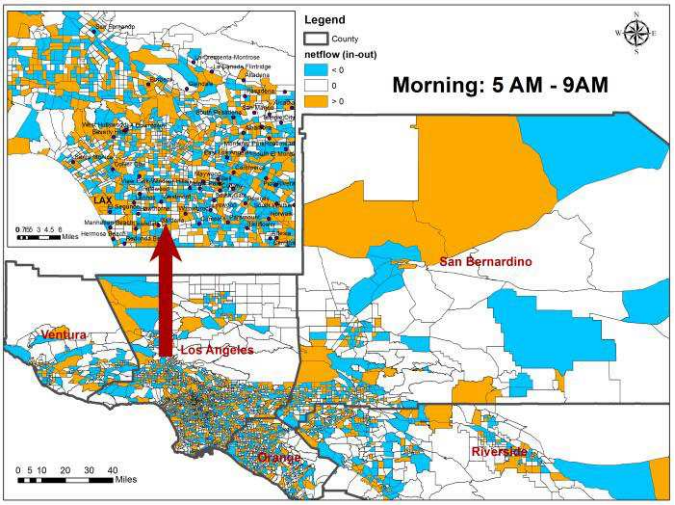

(b)



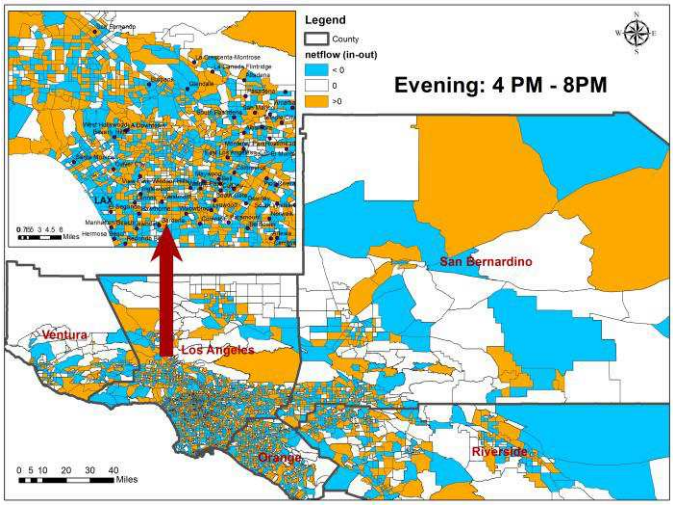

(c)




(d)




(e)




(f)

Figure 3: The spatial distributions of detected OD trips. (a) morning-peak directed pattern; (b) evening peak directed pattern; (c) morning-trip distance distribution; (d) evening-trip distance distribution; (e) morning netflow; (f) evening netflow at the TAZ scale.

### 3.2 OD Trips at the County-level

We spatially aggregated the OD trips from TAZ to the county level. Figure 4 shows the spatial distribution of detected daily OD trips for five counties. Surprisingly, even for such large-scale inter-county and intra-county flow patterns, our results show a perfect rank

matching with the ACS data, which demonstrates that Orange-Los Angeles has the largest inter-county mobility flows, San Bernardino-Los Angeles and Ventura-Los Angeles ranks the second and the third. But most trips occur inside the same county. Moreover, we also detected some regular trips between Riverside and four other counties, which weren't reported in the survey. Such analysis shapes well for the regional-flow patterns and spatial interaction structure, which is beneficial for regional transportation planning.

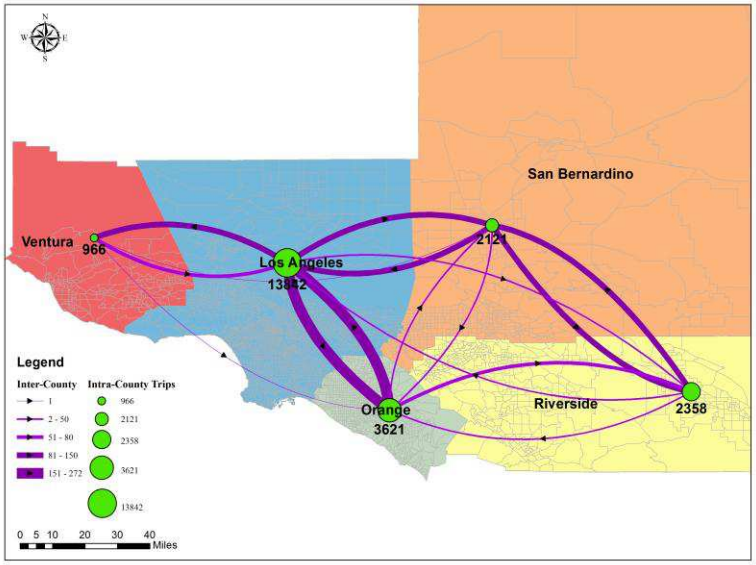



Figure 4: The spatial distribution of detected daily OD trips for five counties.

## 4. Conclusion and Future Work

In this research, we explored the possibility to use large-scale social media data to estimate regional OD trips. Our case study demonstrates that the proposed approach can provide reliable estimates of temporal mobility flows on weekdays compared with the community survey data and also help discover spatiotemporal flow patterns and variations at varying scales. As a next step we would like to investigate how to mine for activity types using the textual parts of tweets.

## References

Calabrese F, Di Lorenzo G., Liu L and Ratti C. 2011. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Computing, 10(4), 36-44.

Guo D, Zhu X, Jin H, Gao P, and Andris C. 2012. Discovering Spatial Patterns in Origin-Destination Mobility Data. Transactions in GIS, 16(3), 411-429.

McNally, MG, 2000. The activity-based approach. In: Hensher, D.A., Button, K.J. (Eds.), Handbook of Transport Modelling. Pergamon, Amsterdam, 53–70.

Liu Y, Kang C, Gao S, Xiao Y, and Tian Y. 2012a. Understanding intra-urban trip patterns from taxi trajectory data. Journal of Geographical Systems, 14(4), 463-483.

Liu Y, Wang F, Xiao Y, and Gao S. 2012b. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, 106(1), 73-87.

Yue Y, Lan T, Yeh AG and Li, QQ, 2014. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. Travel Behaviour and Society, 1(2). 69–78.

# Visualising Emerging Trends of Clusters in a Space-Time Region Using Spatio-Temporal Kernel Regression

T. Nakaya[1], J Haworth[2], T Cheng[2]

[1]Department of Geography, Ritsumeikan University,
56-1 Tojiin-kita-machi, Kita-ku, Kyoto, Japan
Email: nakaya@lt.ritsumei.ac.jp

[2] SpaceTimeLab, Department of Civil, Environment & Geomatics Engineering,
University College London, Gower Street, London WC1E 6BT, UK
Email: {j.haworth; tao.cheng}@ucl.ac.uk

## 1. Introduction

Clusters of crime or disease often appear as spatio-temporal concentrations of cases with emerging trends. In this study, we highlight the visual identification of such emerging trends. Most traditional cluster detection techniques such as space-time scan statistics (STSS) (Kulldorff et al., 1988) evaluate the significance of the incidence (number of cases observed) by comparing it to the expected incidence based on some assumptions, such as the average rate of crime/disease per unit population or area. A cluster is considered to be detected when the observed incidence is significantly higher than the expected (hereafter, we call this the elevated risk model). However, since clusters/outbreaks often evolve from a very low incidence, incidence alone may be an inadequate criterion to judge the detection of the emerging cluster at its early stage (the size of cluster is small but clearly increasing). Alternatively, Tango et al. (2011) proposed to focus on 'emerging trends of incidence' rather than 'magnitude of incidence' for effective cluster detection. They developed a new variant of STSS with an 'outbreak model' (a model with locally linear temporal trend of risk) to capture localized emerging clusters of disease, revealing that the approach detects outbreaks even when the incidence is small.

STSSs, however, makes some strong assumptions, whether using elevated risk or outbreak models. In particular, the geometry of a cluster to be detected must be space-time cylinders and not overlap with another cluster. In addition results of scan statistics with multiple testing adjustments are occasionally too conservative in case that multiple clusters exist. Nakaya and Yano (2010) compared the conventional STSS of the elevated risk model with the space-time kernel density estimation (STKDE), which is used in various fields of geographic information science (e.g. Demšar and Virrantaus, 2010), for detecting space-time clusters of crime. Nakaya and Yano argued that STKDE provides a fuzzier but more detailed description of spatio-temporal distributions of crime clusters compared to the conventional STSS. Thus, we consider a new space-time kernel based non-parametric regression method, GW-LOWESS (Geographically weighted locally weighted scatterplot smoothing), by extending the concept of STKDE to capture localized anomalous emerging trends of crime or disease incidence. The method can be considered as an extension of Tango and his colleagues' outbreak model using a kernel-based approach for a better visual understanding of emerging clusters in a space-time region.

## 2. Method

Consider an aspatial non-parametric temporal model using locally weighted scatterplot smoothing (LOWESS) (Cleveland and Devlin, 1988), which is a classic kernel-based technique for estimating a non-linear smooth bi-variate function, $y_i = f(t_i)$ where $y_i$ and $t_i$ are the incidence and temporal position of the $i$th observation, respectively. The LOWESS model can be described as:

$$\begin{aligned} y_i &= f(t_i) + \varepsilon_i \\ &= \beta_0(t_i) + \beta_1(t_i)t_i + \varepsilon_i, \end{aligned}$$

where $\varepsilon_i$ is the error term and the local coefficients around $t_i$, $\{\beta_0(t_i), \beta_1(t_i)\}$, are obtained by fitting a conventional linear (or quadratic) regression model to a locally weighted subset of the data around the temporal location $t_i$ using a (temporal) kernel function. It is straightforward to evaluate the temporal emerging trend of incidence in this model: if the slope coefficient, $\beta_1(t_i)$, is larger than zero, the epidemic curve, $f$, is increasing.

A related non-parametric regression technique using kernel weighting and local linear model fitting is geographically weighted regression (GWR), which is designed to estimate geographically varying coefficients (Fotheringham et al., 2002). Combining LOWESS and GWR results in GW-LOWESS which is a model of a non-linear temporal functions that depend on the geographical position:

$$\begin{aligned} y_i &= f(t_i \mid u_i, v_i) + \varepsilon_i \\ &= \beta_0(u_i, v_i, t_i) + \beta_1(u_i, v_i, t_i)t_i + \varepsilon_i, \end{aligned}$$

where $(u_i, v_i)$ is the geographic coordinate of observation $i$. To estimate the geographically conditional temporal function (i.e., the geographically localized epidemic curve), we repeatedly fit a local linear model with the temporal variable to a spatio-temporal subset of the data, weighted by a space-time kernel function. The estimation of the local coefficients can be generalized by using the generalized linear modelling (GLM) framework. The estimates are obtained by maximizing the following spatio-temporally weighted log-likelihood function:

$$\{\hat{\beta}_k(u_i, v_i, t_i)\} = \arg\max \sum_j \left\{ \text{log-likelihood}(y_j \mid \hat{y}_j) K_s\left(\frac{u_j - u_i}{h_s}, \frac{v_j - v_i}{h_s}\right) K_t\left(\frac{t_j - t_i}{h_t}\right) \right\}$$

where $Ks$ and $Kt$ are kernel functions for the spatial and temporal domains, respectively, and $h_s$ and $h_t$ are their associated bandwidth parameters.

In the case that crime or disease occurrences are recorded in spatio-temporal aggregated units, a variant of GW-LOWESS based on a Poisson regression scheme is appropriate; specifically,

$$\begin{aligned} y_i &\sim Poisson[\mu_i] \\ \mu_i &= Offset_i \exp\left(\beta_0(u_i, v_i, t_i) + \beta_1(u_i, v_i, t_i)t_i\right), \end{aligned}$$

where $\mu_i$ is the expected count of events and $Offset_i$ is the offset variable of observation $i$. The offset is an adjustment for the size of the observation unit, such as the areal or population at risk. Inferential statistics of this model can be derived by local regression theory. A simple way to assess localized emerging trends is to use the Wald statistic of the temporal coefficient, defined as:

$$z_{slope}(u_i, v_i, t_i) = \beta_1(u_i, v_i, t_i) / SE[\beta_1(u_i, v_i, t_i)] \sim N(0,1).$$

Without multiple testing adjustment, we may consider the trend is significantly increasing at the 5% level when $z_{slope}$ is larger than 1.96.

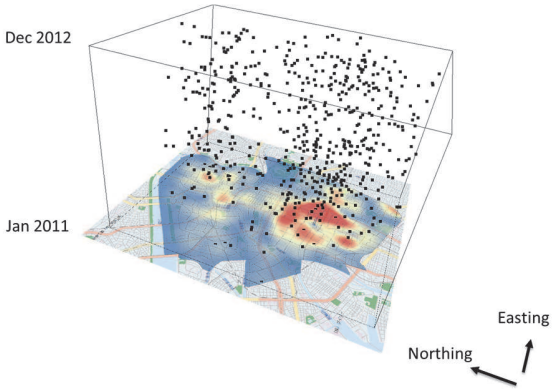



Figure 1 Space-time scatter diagram of crime occurrence.

## 3. Case study

We test the proposed technique on a small dataset of snatch-and-run crime (on-road robbery) incidence in the central part of Osaka City, Japan. The study area consists of 541 tracts (cho-cho aza) and its areal size is 38.0 square kilometres. The total number of reported cases is 611. The temporal resolution is monthly and the data spans from the beginning of 2011 to the end of 2012. Figure 1 shows the spatio-temporal distribution of reported crime in a space-time cube. On the bottom of the cube, the spatial kernel density estimates are plotted (high density regions are coloured in red).

We fit the Poisson regression version of GW-LOWESS to this data using the areal size of tracts, $Area_i$, as the offset:

$$\begin{aligned} y_i &\sim Poisson[\mu_i] \\ \mu_i &= Area_i \times d_i \\ d_i &= \exp\left(\beta_0(u_i, v_i, t_i) + \beta_1(u_i, v_i, t_i)t_i\right), \end{aligned}$$

where $d_i$ is the modelled estimate of crime density at observation $i$. Through a bandwidth selection using AICc (corrected Akaike Information Criterion), values of 1 km and 1 month were chosen for the spatial and temporal bandwidths of the Gaussian kernel, respectively. Using this

kernel weighting, we visualised the gridded crime density estimates and significant emerging trends of crime occurrence (Figures 2) by volume rendering techniques.

Figure 2 (Left) shows the space-time domain having high positive Wald statistics (z = 1.96 and z = 2.50). This indicates that highly positive Wald statistics were present before the space-time domains with high crime density appeared in the south part of the region. Figure 2 (Right) overlays the same space-time contour plots with ray-tracing plots of crime density. In this figure, the high crime density region looks solid coloured in red while lower density regions coloured in blue or green are controlled to have high transparency. Using GW-LOWESS, even lower density clusters in the north part of the study area exhibit high values of the Wald statistic before the relatively high-density domains appear. These results indicate that the emerging trends of crime incidence are a useful description of space-time cluster sequences and can be used for the early detection of potential crime hotspots.

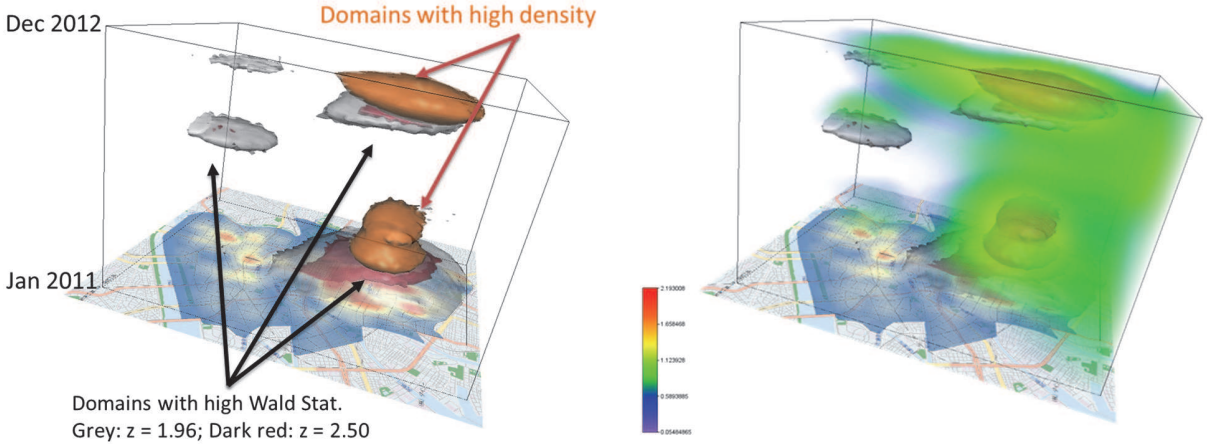



Figure 2 Space-time contours of domains with emerging trends.
(Left) Comparison with contours of high crime densities
(Right) Comparison with the ray-tracing plot of crime densities

## 4. Conclusion

In this study, we proposed a new method called GW-LOWESS as a geographically conditional kernel regression for estimating space-time density as well as emerging trends of crime occurrence in a space-time region. Through statistical testing on the local estimate of the temporal slope of the model, the technique can be used as an early warning system for crime or disease clusters. This finding may have important implications in predictive policing and epidemiology. As the next step, we plan to conduct a comparative analysis of the proposed method with other predictive cluster detection methods such as STSS, by using a larger dataset. Furthermore, we plan several extensions including kernel function weighting only the data in the past (temporally non-symmetrical kernel) for proactive/predictive purposes and a modified testing procedure of local coefficients by randomisation.

## Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24300323 and UK EPSRC EP/J004197/1.

## References

Cleveland, W.S. and Devlin, S.J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83: 596-610.

Demšar, U. and Virrantaus, K. 2010. Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science* 24: 1527-1542.

Fotheringham, A.S., Brunsdon, C., and Charlton, M. 2002. *Geographically Weighted Regression*. Wiley, Sussex.

Kulldorff, M., Athas, W., Feuer, E., Miller, B. and Key, C. 1998. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health* 88: 1377-1380.

Nakaya, T. and Yano, K. 2010. Visualising crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics, *Transactions in GIS* 14: 223-239.

Tango, T., Takahashi, K., and Kohriyama, K. 2011. A space-time scan statistic for detecting emerging outbreaks. *Biometrics* 67: 106–15.

# Validation of Results for Temporal Patterns Derived in STempo

D.J. Peuquet, S. Stehle

GeoVISTA Center, Department of Geography, The Pennsylvania State University, 302 Walker Building, University Park, PA 16802, USA
Email: (Peuquet, Stehle)@psu.edu

## 1. Introduction

Deriving understanding from past real-world events requires finding meaningful patterns that may not be visible within the complex stream of potentially relevant data currently available on the Web. But Web-based reports of events are often highly redundant and frequently contradictory or ambiguous. Moreover, temporal and spatio-temporal patterns of events tend to have hierarchical structures with interactions and feedback effects over multiple spatial and temporal scales. To address this problem, we continue to develop a sophisticated pattern analysis system, called STempo, which provides integrated computational/statistical and visualization tools to help reveal and explore the complex structure of these dynamic associations.

Within STempo we have implemented a statistical/computational technique based on T-pattern analysis for discovering hierarchical associations among events and groups of events. Using data derived from RSS newsfeeds we demonstrate the effectiveness of our technique by comparing patterns of events in Yemen at different time periods and to events in other Middle Eastern countries during Arab Spring.

## 2. Background

There have been significant advances in pattern recognition techniques in the field of data mining, and more specifically, in knowledge discovery from database (KDD) techniques within GIScience. Many procedures have been developed explicitly for temporal and spatio-temporal pattern discovery. These are commonly called sequence analysis procedures and are most frequently used to analyze such things as buying behavior and patterns of bank transactions (Gabadinho et al. 2011). There has also been much activity in analyzing point movement patterns (Joh et al. 2002, Dodge et al. 2012). Nevertheless, existing pattern discovery and pattern analysis techniques tend to focus on the temporal interrelationships among a very limited number of event types. To address the need for computational techniques to help reveal distinct but overlaid temporal relationships in complex space-time data involving many event types, we adapted and extended a statistical pattern discovery technique known as T-pattern analysis (Peuquet 2012).

## 3. Brief Overview of the T-pattern Method

T-pattern analysis, originally proposed by (Magnusson 2000) is designed to detect patterns of associations among event-types, where a specific sequence of event types recurs more often than is likely by chance. Data is organized in a series of timelines, where each timeline consists of the ordered sequence of times when events of a given event type occurred. There is thus one timeline for each event type. This ordering also allows redundancies to be eliminated, as only the first of multiple events with the same type and timestamp will be retained in the analysis. Events may have duration, however. These are represented as a sequence of consecutive time stamps.

In order to detect patterns, two timelines of occurrences for two different event types are compared to find whether the temporal distances between co-occurrences are random or not with

respect to a user-specified level of statistical significance (p-value). This test, called the Critical Interval (CI) test, is based on the null hypothesis that pairs of events (e.g., A and B) are temporally independent of each other against the alternative that the real time differences, $d$, between them are relatively invariant. Further, if a type A event occurs at time $t_i$, the CI test checks whether there is an event type B occurring within a time interval of $[t_i+d_1, t_i+d_2]$ (where $d_2>d_1\geq 0$) more often than we would expect by chance. Since there can be no assumption about which event type tends to occur before the other, the CI test checks both BA and AB occurrences. The temporal distance given in the CI, as well as the use of a probabilistic approach accommodates non-stationarity of event patterns and provides a degree of fuzziness in pattern identification.

The pattern detection process proceeds in a bottom-up fashion by comparing each event timeline with all other event timelines, including timelines that had been combined from previously detected groupings of event types. Resulting patterns may be complex hierarchies comprised of multiple event types. T-pattern analysis only identifies as patterns those sequences of event types that consistently recur within a given span of time, along with expected temporal distance (the CI) between each of those events. There is no implication of causality.

## 4. Validation Analysis Design

Given this probabilistic approach, we would not expect T-pattern analysis to find significant patterns in randomized data. We thus designed a more complex validation analysis to compare results using subsets of real-world data.

Patterns of real-world events change and evolve over time. Also, patterns of events can be expected to recur in places that are similar. Results from T-pattern analysis should therefore reflect this. To test our expectations of how T-pattern analysis will perform on real-world event data, we select a specific pattern discovered to be significant by T-Pattern analysis and test whether this pattern is found in other spatial and temporal contexts. Where we do find recurrence, we examine potential changes in CI values.

While T-pattern analysis can be applied to any domain of real-world events and at any spatio-temporal scale (the spread of disease, climate change, etc.), our validation analysis focuses on events in the Middle East. In particular, the events of Arab Spring embody new social and political patterns that cannot always be assumed based on *a priori* canonical examples. But not all countries in the Middle East have undergone political change. Some remain stable. We therefore expect to find variations in patterns of social and political events between countries and variations over time for a country undergoing change.

### 4.1 The Data

We captured RSS news feed data and subsequently encoded event types for each data observation from the article titles using the TABARI and CAMEO categorization tools (Schrodt et al. 2008). We found that some built-in CAMEO event types (i.e., ‘reject’ and ‘disapprove.’) are overly vague for the purposes of finding meaningful patterns. We therefore also implemented a post-processing program to expand the event ontology and further specify the frequently generic event types into more specific categories. Because event location is often not mentioned in news report titles, we used part-of-speech tagging on the body of the story to determine geographic location. HTML tags contained within the source code of web-based news reports were used to extract the body of the story.

### 4.2 Comparison of Patterns among Times and Places

We began by running the T-pattern discovery process on event data for only the country of Yemen from February, 2011 through March 2012 as our basis for comparison. Figure 1 shows one pattern revealed by T-pattern analysis that is seemingly indicative of the socio-political instability in Yemen at that time. Event types are represented by numerical codes, and event transition is given as the CI in number of days.

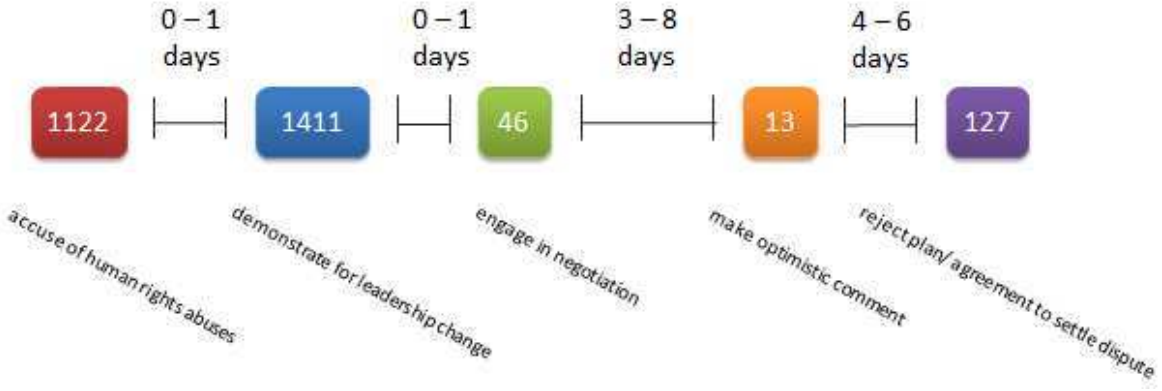



Figure 1: A temporal pattern of diplomatic cooperation discovered in Yemen

An accusation of human rights abuses (event code 1122) is followed within one day by a demonstration for a change in leadership (1411), followed within another day by engagement in negotiation (46), followed between three and eight days by an optimistic comment being made (13), concluding four to six days later with rejection of the plan that would have settled the dispute (127).

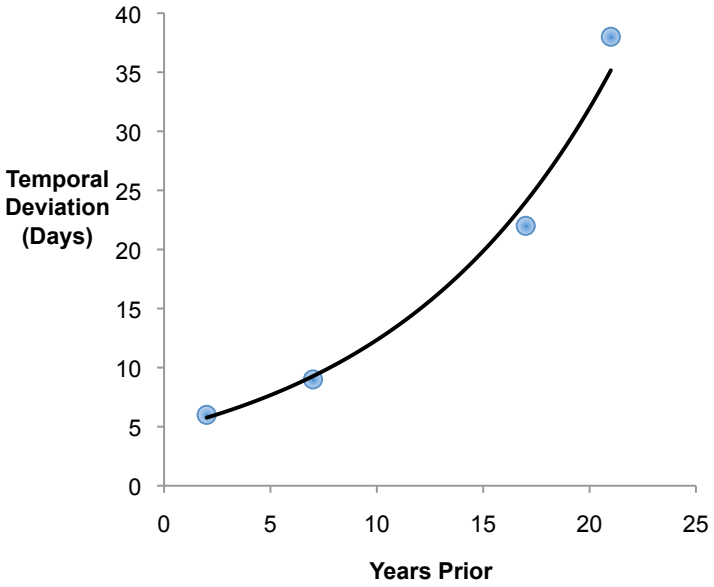



Figure 2: Temporal variation of selected pattern in Yemen

We then ran additional T-pattern analyses on Yemen for four temporal intervals that were politically unstable times for that country (2009-10, 2004, 1994, and 1989-91). We found this same pattern, but with differing CI values. As shown in Figure 2, CI values (in number of days) increase with increasing temporal distance (in years) from the 2011-2012 temporal interval. This would indicate an evolving situation in that country.

For our continuing validation analysis, we selected Syria, Libya and Egypt as countries with similar socio-political instability, and Bahrain, Jordan and United Arab Emirates as countries that were stable during this period.

## REFERENCES

Dodge, S., Laube, P. and Weibel, R., 2012. Movement similarity assessment using symbolic representation of trajectories. *International Journal of Gegraphical Information Science*, 26(9), pp. 1563-1588.

Gabadinho, A., Ritschard, G., Studer, M. and Müller, N.S., 2011. Extracting and rendering representative sequences. In *Knowledge Discovery, Knowledge Engineering and Knowedge Management*, J.L.G.D. A. Fred, K. Liu and J. Filipe (Ed.), pp. 94-106 Berlin: Springer-Verlag.

Joh, C.-H., Arentze, T.A. and Timmermans, H.J.P., 2002. Activity Pattern Similarity: A Multidimensional Sequence Alignment Method. *Transportation Research Part B*, 36, pp. 385-403.

Magnusson, M.S., 2000. Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers*, 32(2), pp. 93-110.

Peuquet, D.J., 2012. A method for discovery and analysis of temporal patterns in complex space-time event data. In *Seventh International Conference on Geographic Information Science (GIScience 2012)* Columbus, OH.

Schrodt, P.A., Yilmaz, Ö., Gerner, D.J. and Hermreck, D., 2008. The CAMEO (Conflict and Mediation Event Observations) actor coding framework. In *International Studies Association Annual Meeting* San Francisco.

# What is Field and Object Information?

Werner Kuhn

Department of Geography and Center for Spatial Studies
University of California, Santa Barbara CA, 93101
Email: kuhn@geog.ucsb.edu

## Abstract

The distinction between field and object models is well established in Geographic Information Science. Yet, despite the vast literature on the subject (see, for example, (Couclelis, 1992; Galton, 2004; Goodchild, Yuan, & Cova, 2007; Worboys & Duckham, 2004)), no commonly available definitions specify precisely what information field or object models provide. Experts disagree not just on the details, but on basic questions, such as whether fields are defined over discrete or continuous spaces or both, whether objects require boundaries or not, and whether fields and objects are something in the world, in our minds, or in information systems. Depending on the positions taken on some of these questions, population densities can be modeled as fields or not and mountains as objects or not, to give just two examples.

This paper specifies fields and objects in terms of core queries to be answered by any of their implementations. All object and field data types should be derivable from the proposed specifications, demonstrating thereby how implementations answer the core queries. The specifications are part of a collection of formally specified classes of spatial data types, pushing the core concepts of spatial information proposed in (Kuhn, 2012) to the level of a high-level language for GIS. This language will allow users to ask questions about spatial phenomena without the need to know representational or algorithmic details. The main difference to previous attempts at organizing GIS functionality is that this one starts from patterns for questions rather than from the complexity of answering methods in the form of GIS commands and data models.

## 1. Motivation

In an attempt to define formally what GIScience is about, and adopting a wider spatial (as opposed to just geographic) scope, I have proposed a set of core concepts of spatial information (Kuhn, 2012). Fields and objects are, not surprisingly, two key entries on that list and would probably be on most other lists coming from GIScientists. However, when attempting to specify them in more detail, one finds disagreements on their exact nature.

The main goal of this short paper is, therefore, to advance the discussion on what characterizes fields and objects. They are seen here as conceptualizations of real-world phenomena, not to be confused with the phenomena themselves, nor with their computer models. The phenomena can always be conceptualized differently (for example, as processes) and the implementations can take many forms. Temperature, for example, is not a field per se, nor is a temperature modeled implemented as a field, but temperature can be conceptualized as a field and implemented by raster data, point data with interpolation functions, stored functions, or otherwise.

The remainder of the paper proposes a specification for fields (section 2), one for objects (section 3), and a conclusion on how such high-level specifications help improve the usability of

spatial information technologies (section 4). The appendix provides the current state of the field and object specifications, with a link to their evolving and executable online versions.

## 2. Fields

Fields capture phenomena that can be described by properties with unique values everywhere in a space of interest (Goodchild et al., 2007). Such so-called *intensive* properties can be represented by scalars, vectors, spinors, or tensors, though GIScience deals primarily with the first two of these. The values can result from measurements or computations, thus including derived quantities like probabilities and densities. Typical geographic field examples are temperature and wind fields. A temperature field provides a scalar value and a wind field a velocity vector (speed and direction) for each position.

Mathematically, fields are characterized by *continuous functions* from positions to values, meaning that small changes in positions lead to small changes in values. Note that positions as well as values can be discrete, while still allowing for continuous functions between them (Rosenfeld, 1986). In practice, the continuity condition on the function is sometimes ignored, retaining just a functional relationship between positions and values. Land cover or land ownership are examples of phenomena requiring this broader definition, as their values do not change smoothly.

The *domain* of a field function typically contains a subset of all possible positions in the chosen reference system. For example, temperature and wind fields may be given for a country or continent only. The domain can be a union of several such subspaces. Domains may be described by their boundaries.

The *positions* can be in spaces with any number of dimensions, though the spaces of human experience always have three dimensions or less. For example, temperatures on the surface of the earth require positions with two, temperatures inside a building require positions with three dimensions. Positions can be spatio-temporal, though time often gets separated from space and modeled as snapshots.

The core *operations* on fields ask for or change the values at positions. In addition, the local, focal, and zonal operations of Tomlin's map algebra (Tomlin, 1983) can be lifted from their raster implementation origins to the field concept (Worboys & Duckham, 2004). They derive new field values for each position, based on the values at that position, in its neighborhood, or in a zone containing it.

An algebraic specification of this core idea of fields is given in Appendix A. It largely follows (Worboys & Duckham, 2004) and has been developed in the same spirit and style as the one in (Camara et al., 2014), but shows different results from the latter, because it specifies the field concept, not its implementation.

## 3. Objects

Objects capture phenomena that carry *identity* and are *bounded*, i.e. contained in a portion of space. Geographic examples are buildings or lakes. Through their identity, objects can be asked for their properties and relationships and their changes can be tracked over time.

As the properties pertain to whole objects, they are called *extensive*. The properties and relationships can be spatial, temporal, or thematic. Some of them form concepts themselves and are therefore not further specified here. An example is the core concept of *location,* which is considered as a collection of spatial relations applicable to objects, field zones, networks, and events.

In GIScience, objects are often seen as necessarily having *boundaries*, sometimes even crisp ones. Yet, the extension of the concept from common sense manipulable objects to larger (for example, geographic) scales suggests that objects do not need to have a boundary, though they are always bounded. Many natural objects, such as forests or beaches, have transition zones between what belongs to them and what does not (Burrough, 1996). The insistence on boundaries is a leftover from early object (vector) data produced by digitizing maps and images.

*Parts* of objects can themselves be treated as objects, and complex objects get aggregated from simpler ones. Part-whole relations are central to object models and have spatial and temporal aspects themselves. For example, a condominium can be located within a building and may only exist for part of the lifetime of the building, which may itself be part of a building block for some time. Such partonomic hierarchies of objects and their parts often extend over several levels.

*Feature*s are parts of surfaces of objects and can be considered special cases of objects when they are given their own identity. For example, while lakes may be seen as three-dimensional objects, they are also often treated as (independent) two-dimensional parts of the earth's surface. The distinction between features and objects is not verbalized in many languages, so that it makes sense to ignore the difference and treat both as objects.

This general notion of object, formally specified in Appendix B, is simple but extremely versatile. In addition to prototypical examples like buildings, it also covers, for example, mountains, packets of cold water in an ocean, or ash clouds in the atmosphere, as well as objects in non-geographic spaces.

## 4. Conclusions

Precise specifications of spatial information enable more usable spatial computing software, which in turn allows for a broader exploitation of location in science and decision-making. With this goal in mind, I am developing my core concepts of spatial information (Kuhn, 2012) into a high-level language for asking and answering spatial questions. The preliminary results reported here for the concepts of field and object suggest that fields are best characterized by continuous (or sometimes non-continuous) functions and objects by identity. This contrasts with the frequent assumptions that fields require continuous spaces or continuous values or both and that objects require boundaries. It also draws the important distinction between reality, its conceptualizations, and the implementations of these. The validation of the proposed specifications by instantiating GIS commands to the specified operations is currently under way.

## Acknowledgements

The support from the UCSB Executive Vice Chancellor for the Center for Spatial Studies and from Jack and Laura Dangermond for GIScience students working as summer interns on this project is gratefully acknowledged.

## References

Burrough, P.A. and Frank, A.U. (1996). *Geographic objects with indeterminate boundaries*. London: Taylor&Francis.

Camara, G., Egenhofer, M. J., Ferreira, K., Andrade, P., Queiroz, G., Sanchez, A., Jones, J., Vinhas, L. (2014). Fields as a Generic Data Type for Big Spatial Data. M. Duckham et al. (Eds.): GIScience 2014, LNCS 8728 (in press).

Couclelis, H. (1992). People manipulate objects (but cultivate fields): Beyond the raster-vector debate in GIS. In A. U. Frank, I. Campari, & U. Formentini (Eds.), *Theories and methods of spatio-temporal reasoning in geographic space* (pp. 65–77). Berlin: Springer-Verlag.

Galton, A. (2004). Fields and Objects in Space, Time, and Space-time. *Spatial Cognition & Computation*, *4*(1), 39-68.

Goodchild, M. F., Yuan, M., & Cova, T. J. (2007). Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, *21*(3), 239–260.

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, *26*(12: Special Issue in honor of Michael Goodchild), 2267–2276.

Rosenfeld, A. (1986). 'Continuous' functions on digital pictures. *Pattern Recognition Letters*, *4*(3), 177–184.

Tomlin, C. D. (1983). A Map Algebra. In *Harvard Computer Graphics Conference* (unpublished handout). Cambridge, MA: Harvard University, Graduate School of Design.

Worboys, M. F., & Duckham, M. (2004). *GIS: A Computing Perspective, Second Edition* (p. 448). CRC Press.

## Appendix: Formal Specifications

The specifications for the field and object concepts are given in the form of Haskell type classes, i.e. algebraic specifications of the operations to be provided by all field or object types (see http://haskell.org for tutorials and compilers). The code can be viewed, used, and discussed at https://www.fpcomplete.com/user/kuhn/core-concepts-of-spatial-information.

### A. Fields

The `FIELDS` class specifies that all field types need an operation `get` to return the field value at a position and an operation `set` to change the value at a position. The result of `domain` can, for example, be an interval, a rectangle, two corner points, a convex hull or a boundary. The `neighborhood` and `zone` functions return a set of positions or a boundary. The map algebra functions `local`, `focal`, and `zonal` compute a field with new values based on the previous values at each position, its neighborhood, or in a zone containing it.

```
class FIELDS field pos val where
  get :: field pos val -> pos -> val
  set :: field pos val -> pos -> val -> field pos val
  domain :: field pos val -> Geometry
  neighborhood :: field pos val -> pos -> Geometry
  zone :: field pos val -> pos -> Geometry
  local :: field pos val -> (val -> val') -> field pos val'
  focal :: field pos val -> (pos -> val') -> field pos val'
  zonal :: field pos val -> (pos -> val') -> field pos val'
```

### B. Objects

The class `OBJECTS` of all object types is derived from the class `Eq` of types that can be compared for equality, which provides identity. Objects can be asked for their bounds (a geometry within which they lie). The operation `get` returns the value of a property defined by `obj -> val` and `is` determines whether two objects are in the relation `obj -> obj -> Bool`.

```
class Eq obj => OBJECTS obj where
  bounds :: obj -> Geometry
  get :: (obj -> val) -> obj -> val
  is :: (obj -> obj -> Bool) -> obj -> obj -> Bool
  get property = property
  is relation = relation
```

# MapReduce Principle for Spatial Data

F.-B. Mocnik

Vienna University of Technology, Gußhausstraße 27-29, 1040 Vienna, Austria
mocnik@geoinfo.tuwien.ac.at

## 1. Introduction

The amount of data accessible and used for spatial reasoning is rapidly increasing, in particular because efforts have been undertaken to make relations between data sets explicit, e. g. by the approach of linked data (Bizer, Heath and Berners-Lee 2009). It becomes increasingly hard to process data in acceptable time and complex reasoning is in many cases hardly possible. Data exhibiting these characteristics is called *big data* (Akerkar 2013, Lynch 2008, Snijders, Matzat and Reips 2012) and is relevant for GIScience (Adams, Brodaric, Corcho et al. 2012).

To address the processing of large amounts of data, parallelism becomes more important in order to use full processing power on multi-core and multi-processor computers as well as on computer clusters (Cannataro, Talia and Srimani 2002). MapReduce has been proven to be a successful approach that follows the paradigm of parallelism (Dean and Ghemawat 2004).

Libraries for applying MapReduce to spatial problems have already been implemented (Cary, Sun, Hristidis et al. 2009, Chen, Wang and Shang 2008, Eldawy and Mokbel 2013, Wang, Han, Tu et al. 2010). Research has been done on different aspects, like balancing the work between different cores (Chen, Chen and Zang 2010) and how to use spatial joins (Zhang, Han, Liu et al. 2009) and spatial indexes (Zhong, Han, Zhang et al. 2012) with MapReduce.

In this paper we discuss when and how MapReduce can and when it cannot be applied to spatial problems, and how to modify the problem in the latter case such that MapReduce can be applied to at least parts of the problem.

## 2. MapReduce Principle and Spatial Data

MapReduce is a principle to parallelize computations on large data sets: (i) the given computational problem $P$ can be partitioned into several smaller ones $P_i$, (ii) the problems $P_i$ can be computed in parallel (this step is called mapping) and finally, (iii) the results are combined using a reduction function (Dean and Ghemawat 2004). This principle works as long as two major requirements are met: first, the overall problem can be partitioned, and secondly, the overall result can inexpensively be computed based on the computations' results of the partial problems. Many problems do meet these requirements but others do not.

Spatial data refers to space. In the following, we only consider spatial data which can be associated to spatial regions. For example, to a region $U$ we can associate the number of trees, objects depicted in a map or information about bus stops. The data $\mathcal{F}(U)$ assigned to a region $U$ can be mapped to some data $\mathcal{G}(U)$ which is the result of the computational problem.[1]

Applying MapReduce to a *spatial* problem $P\colon \mathcal{F}(U) \mapsto \mathcal{G}(U)$ suggests itself to take advantage of its spatial structure: a partition of space $U = \bigcup_i U_i$ leads to data sets $\mathcal{F}(U_i)$ which can be used to formulate smaller problems $P_i\colon \mathcal{F}(U_i) \mapsto \mathcal{G}(U_i)$. (In fact, the mapping relation should be the same for all $P_i$. Thus, we omit the index and simply write $P'\colon \mathcal{F}(U_i) \mapsto \mathcal{G}(U_i)$.) In the reduction step, the resulting datasets are combined to the result of the main problem by computing $\mathcal{G}(U) = \bigoplus_i \mathcal{G}(U_i)$.[2]

[1] The notation $\mathcal{F}(U)$ is based on the mathematical concept of a presheaf.

[2] This notation of the MapReduce principle mentions the used indexes only implicitly: the keys in before the mapping correspond to the indices $i$, and after the mapping step there exists only one index, allowing to omit it.

## 3. Conditions for MapReduce on Spatial Problems

Computations can in general not be divided into smaller computations which can be processed in parallel (Bernstein 1966). For applying MapReduce to a spatial problem $P\colon \mathcal{F}(U) \mapsto \mathcal{G}(U)$, the following conditions are sufficient:

**Condition (C1)** There exists a partition $U = \bigcup_i U_i$, data sets $\mathcal{F}(U_i)$ associated to each $U_i$ and a mapping $P'\colon \mathcal{F}(U_i) \mapsto \mathcal{G}(U_i)$.

**Condition (C2)** There exists an operation $\oplus$ such that $\mathcal{G}(U) = \bigoplus_i \mathcal{G}(U_i)$ and $\oplus$ is inexpensive to compute.

Even if the conditions are not met, one may be able to conclude the solution $P(\mathcal{F}(U))$ by only using the $\mathcal{G}(U_i)$ and some additional information like the $U_i$ and/or additional spatial data. Thus, the conditions are not necessary for applying MapReduce. On the other side, if both conditions are met, we are able to apply MapReduce because condition (C1) implies the partition (induced by its spatial structure) of step (i) as well as the existence of the problems $P_i$ of step (ii). Condition (C2) ensures that the solutions of the problems $P_i$ can be combined to the solution of the original problem $P$ because the following holds:

$$P(\mathcal{F}(U)) = \mathcal{G}(U) \overset{\text{(C2)}}{=} \bigoplus_i \mathcal{G}(U_i) \overset{\text{(C1)}}{=} \bigoplus_i P'(\mathcal{F}(U_i))$$

This proves the conditions (C1) and (C2) to be sufficient for applying MapReduce.

Consider the following examples:

*Example.* (1) *Number of objects of a certain kind.* $\mathcal{F}(U)$ is a collection of objects that are located in $U$. Consider the problem $P\colon \mathcal{F}(U) \mapsto \mathcal{G}(U)$ of counting the objects of a certain kind $\kappa$, e. g. trees or street crossings, i. e. $\mathcal{G}(U)$ is a number.

For applying MapReduce, we choose an arbitrary partition $U = \bigcup_i U_i$ and define $\mathcal{F}(U_i) \subset \mathcal{F}(U)$ to be the subsets of all objects located in $U_i$. The problems $P_i\colon \mathcal{F}(U_i) \mapsto \mathcal{G}(U_i)$ map a collection of objects to the number of objects of kind $\kappa$. As every object in $\mathcal{F}(U)$ is contained in one (and only one) $\mathcal{F}(U_i)$, we can define $\oplus$ as the addition:

$$\mathcal{G}(U) = \sum_i \mathcal{G}(U_i) =: \bigoplus_i \mathcal{G}(U_i).$$

(2) *Statistical properties.* Whenever a statistical property is a quantity $q$ set into proportion to the area as statistical population (in this case the area of $U$), one may compute $q$ using MapReduce and then divide the result after the reduction step (iii) by the area. Observe that the division is an additional step to the MapReduce principle.

## 4. Reformulation of Problems for MapReduce

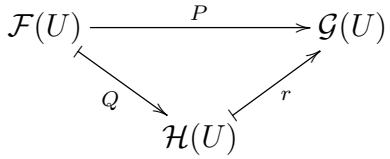
If a problem does not meet condition (C2), one may modify it such that the property is met. In the following, we will discuss an example and its exemplary modification to finally formulate a rule which can characterize if a modification is optimal.

*Example.* (3) Let $\mathcal{F}(U)$ be time table information about when and which bus line does stop at which bus stop $s \in U$, and $\mathcal{F}(U_i) \subset \mathcal{F}(U)$ the corresponding subsets with all bus stops located in $U_i$. Assume that our computational problem $P\colon \mathcal{F}(U) \to \mathcal{G}(U)$ is to compute how many different bus lines meet pairwise, i. e. the number of (unordered) pairs $(b_1, b_2)$ of bus lines where $b_1$ and $b_2$ meet at least at one stop.

The problem $P$ cannot directly be solved by applying MapReduce because for combining the results of smaller problems in the reduction step (iii), the number of pairs for each $U_i$ is not

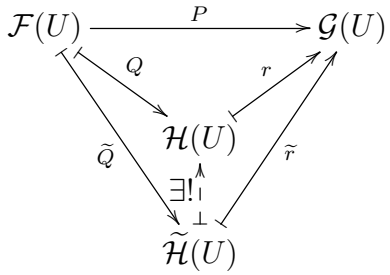
sufficient for not counting a pair several times if it is element of $\mathcal{F}(U_i)$ and $\mathcal{F}(U_j)$ with $i \neq j$, i. e. if two bus lines meet more than once.

To use MapReduce in spite of this we define (for an arbitrary partition $U = \bigcup_i U_i$) smaller problems $Q' \colon \mathcal{F}(U_i) \mapsto \mathcal{H}(U_i)$ which compute the pairs for $U_i$. Thus, we are able to compute all pairs $\mathcal{H}(U) = \bigcup_i \mathcal{H}(U_i)$ as the union of all sets of pairs. (Every pair occurs at most once in the union.) After this, we apply a function $r \colon \mathcal{H}(U) \mapsto \mathcal{G}(U)$ which counts the number of pairs. This function $r$ is not part of MapReduce but is needed to compute the problem $P$.




This principle can also be used in general: a function is executed after the MapReduce step, and/or one before. Commonly, there is more than just one possibility to choose these functions. To determine which choice is optimal for computational purposes, a detailed analysis of the algorithm regarding its performance on a specific computer system would be necessary. However, we can formulate a formal property that ensures that as much as possible of the problem can be computed in parallel.

Assume the situation of MapReduce with a problem $Q'$ (which can be computed using MapReduce) and a function $r$ executed afterwards as well as alternative choices $\widetilde{Q}$ and $\widetilde{r}$:




If the choice of $Q$ and $r$ meets the following criterion, it is ensured that $Q$ computes at least as much as $\widetilde{Q}$, and thus as much as possible is computed in parallel:

**Optimality Criterion (OC)** For every choice $\widetilde{Q}$ and $\widetilde{r}$ with $P = \widetilde{r} \circ \widetilde{Q}$ and $\widetilde{Q}$ computable using MapReduce, there exists one (and only one) map $i \colon \widetilde{\mathcal{H}}(U) \mapsto \mathcal{H}(U)$.

In general there may not exist any $Q$ and $r$ that meets this optimality criterion. If however the optimality criterion is met for a certain $Q$ and $r$ as well as for $\widetilde{Q}$ and $\widetilde{r}$, there exists one (and only one) map $i \colon \widetilde{\mathcal{H}}(U) \mapsto \mathcal{H}(U)$ and one (and only one) map $\widetilde{i} \colon \mathcal{H}(U) \mapsto \widetilde{\mathcal{H}}(U)$. Thus, $i$ is an isomorphism with $\widetilde{i}$ as its inverse.

In the example given above, we may e. g. define $\widetilde{\mathcal{H}}(U)$ as a collection of unique identifiers which correspond to each possible tuple of bus lines. The functions $i$ and $\widetilde{i}$ correspond to the translation between the pairs and the unique identifiers.

## 5. Conclusion

MapReduce cannot be applied to some spatial problems whilst it can be applied to others. We did formulate two sufficient formal conditions for being able to apply MapReduce. If these conditions are not met and MapReduce cannot be applied, one may be able to reformulate the problem such that MapReduce can be applied to a part of it. By the given optimality criterion, we are able to ensure that as much as possible of the computational problem is parallelized. This could be especially useful for formal verification of performance properties and as break condition of automatic parallelization algorithms.

## References


Adams B, Brodaric B, Corcho O et al., eds., 2012, *Proceedings of the Workshop on GIScience in the Big Data Age. In conjunction with the 7th International Conference on Geographic Information Science (GIScience) 2012*.

Akerkar R, ed., 2013, *Big Data Computing*. Chapman and Hall/CRC, London.

Bernstein AJ, 1966, Analysis of Programs for Parallel Processing. *IEEE Transactions on Electronic Computers*, EC-15(5):757–763.

Bizer C, Heath T and Berners-Lee T, 2009, Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Cannataro M, Talia D and Srimani PK, 2002, Parallel data intensive computing in scientific and commercial applications. *Parallel Computing*, 28:673–704.

Cary A, Sun Z, Hristidis V et al., 2009, Experiences on Processing Spatial Data with MapReduce. In: *Proceedings of the 21st International Conference on Scientific and Statistical Database Management (SSDBM) 2009*, 302–319.

Chen Q, Wang L and Shang Z, 2008, MRGIS: A MapReduce-enabled High Performance Workflow System for GIS. In: *Proceedings of the 4th International Conference on eScience 2008*, 646–651.

Chen R, Chen H and Zang B, 2010, Tiled-MapReduce: Optimizing Resource Usages of Data-parallel Applications on Multicore with Tiling. In: *Proceedings of the 19th International Conference on Parallel Architectures and Compilation Techniques (PACT) 2010*, 523–534.

Dean J and Ghemawat S, 2004, MapReduce: Simplified Data Processing on Large Clusters. In: *Proceedings of the 6th Symposium on Operation Systems, Design and Implementation (OSDI) 2004*.

Eldawy A and Mokbel MF, 2013, A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data. In: *Proceedings of the 39th International Conference on Very Large Data Bases (VLDB) 2013*.

Lynch CA, 2008, Big data: How do your data grow? *Nature*, 455(7209):28–29.

Snijders C, Matzat U and Reips UD, 2012, Big Data: Big Gaps of Knowledge in the Field of Internet Science. *International Journal of Internet Science*, 7(1):1–5.

Wang K, Han J, Tu B et al., 2010, Accdelerating Spatial Data Processing with MapReduce. In: *Proceedings of the 16th International Conference on Parallel and Distributed Systems (ICPADS) 2010*, 229–236.

Zhang S, Han J, Liu Z et al., 2009, SJMR: Parallelizing Spatial Join with MapReduce on Clusters. In: *Proceedings of the International Conference on Cluster Computing and Workshops (CLUSTER) 2009*.

Zhong Y, Han J, Zhang T et al., 2012, Towards parallel spatial query processing for big spatial data. In: *Proceedings of the 26th International Parallel and Distributed Processing Symposium Workshop and PhD Forum (IPDPSW) 2012*, 2085–2094.

# An Algorithm for Random Trajectory Generation Between Two Endpoints, Honoring Time and Speed Constraints

G. Technitis, R. Weibel

Department of Geography, University of Zurich (UZH),
Winterthurerstrasse 190, CH-8057, Zurich
Email: {georgios.technitis | robert.weibel}@geo.uzh.ch

## 1. Introduction

In order to test hypotheses regarding possible effects of stimuli on the movement of an animal, researchers frequently use various forms of random walk (RW) as a reference movement model (Turchin 1998). Using the starting point (origin) and the diffusion coefficient of the walk as an input, multiple instances of RWs may then be used to create a space utilization surface, a valuable tool for the behavioral ecologists.

However, in cases where movement has to be simulated between two given points (i.e. an origin and a destination) in a given limited timeframe — for example simulating potential paths between two stop-overs in continental scale bird migration — simple RWs fail to deliver the desired result. Extensions, such as RW with correlation, bias (or drift), self-avoidance abilities, etc. have thus been introduced. Bartumeus et al. (2005) used multiple distributions of correlation for increasing the 'search efficiency' of the walk, and Codling et al. (2008, 2010) attempted to quantitatively parameterize it. In a more qualitative approach, Fronhofer et al. (2013) try to cluster and express the behavior of the animal with a biased correlated random walk and 'area restricted search strategies'. In order to meet the constraint of reaching the destination point, a common solution is to introduce a global bias to 'force' the RW to the endpoint, thus resulting in a highly unrealistic mode of movement which, at the same time, neglects the temporal dimension (i.e. the available time of total travel). Thus, RW models can only be made sufficiently efficient at the expense of introducing excessive bias, thus sacrificing randomness.

Space-time prisms and the Brownian bridges movement model, which will be discussed below, account for the time of total travel, though both methods focus on calculating areas and surfaces rather than individual trajectories. Thus, the use for hypothesis testing of individual movement patterns is limited.

We propose an algorithm that combines concepts from random walks, space-time prisms and alibi queries, and is capable of efficiently generating trajectories between a given origin and a destination, with the least bias possible. Since it is implemented both on the plane and the sphere it is suitable for simulating intercontinental movements such as those of migrating birds. The algorithm is catering to applications in animal ecology. It is not intended to simulate human navigation and wayfinding (e.g. on networks), though it accounts for physical limitations, such as maximum speed and movement time, and provides the user with either single or multiple trajectories as a result. The single trajectory can be used as an (unbiased) null model to test hypotheses about movement stimuli (bias), while the multiple trajectories may be used to create a probability density surface for comparison against established methods, most importantly Brownian bridges.

## 2. Background

Space-time prisms (STP) allow to define an envelope of accessibility given a specific time budget and maximum speed (Hägerstrand 1970), resulting in a potential path space (in a 3D

cube) or its 2D projection, the potential path area (PPA) as shown in Figure 1. One of the methods related to our study was based on the STP and developed by Kuijpers et al. (2011) in an attempt to identify whether two spatio-temporal moving objects, given a maximum possible speed, could have physically met (solving the alibi query). The reasoning of the solution is simple and concrete, though the result is a surface, not an individual trajectory.

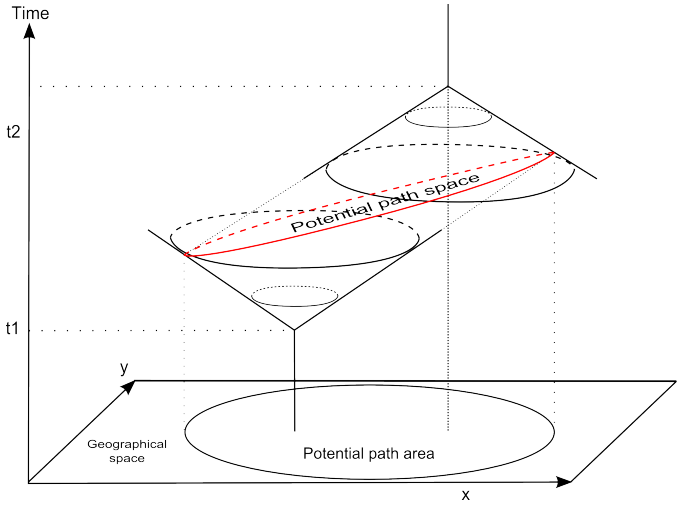



Figure 1. The space-time prism (STP) principle (adapted from Miller 1991 and Kuipers 2011)

The Brownian bridge (BB) movement model attempts to account for both origin and destination, considering the time of total travel. Since the BB was introduced in the ecology community, it has been used extensively for defining the home range of species (Bullard 1991, Benhamou 2011), by calculating the space utilization distribution. In the example of Figure 2, which was created using the ArcMET BBMM tool by Jake Wall, we can see that this method generates a probability surface.

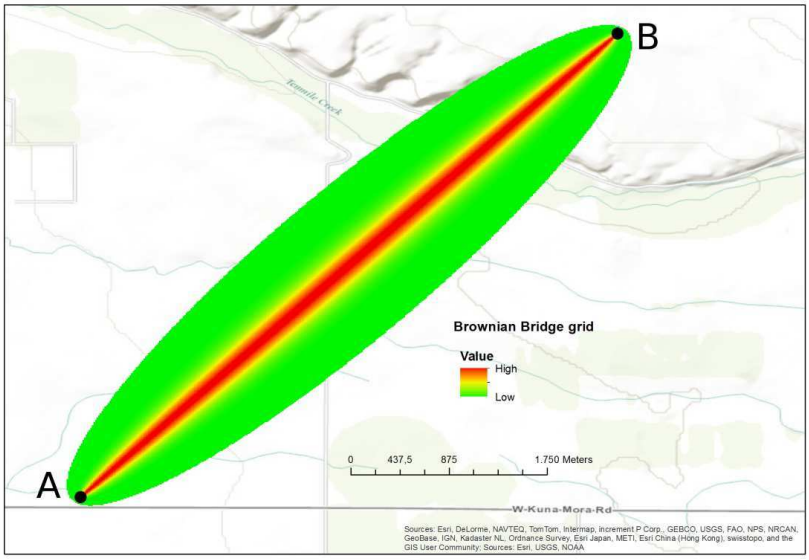



Figure 2. An example of a 0.99 probability surface generated between two points using Brownian bridges.

Recently, Song and Miller (2014) have presented a possible combination of the two approaches, with the definite advantage that the BB probability can be truncated using the STP, excluding areas that cannot be reached in the given timeframe. However, this method focuses on generating a truncated BB surface rather than a single trajectory, and even if a

trajectory is sampled from the resulting visit probability surface, no analytical solution exists to compute the STP on the globe. This deficiency becomes apparent in applications requiring distance calculations over large geographic distances.

## 3. Algorithm

We now describe an algorithm that generates trajectories between two given points $A$ (origin) and $B$ (destination) as randomly placed walks with as little bias as possible. We accomplish this by integrating concepts of the STP model in random walk generation.

Based on the reasoning of the STP model a mobile object can move, at every time-step, as far as its maximum speed ($V_{max}$) permits. At the same time, $n$ time-steps before the end of the trajectory, the object should be within the reach of the destination $B$. The candidate area where an intermediate point of the simulated trajectory can be placed is called the potential *point* area (PpA), not to be confused with the potential *path* area (PPA). The PpA is the intersection of two circles, in our case, the circles $C_A(A,r_A)$ and $C_B(B,r_B)$. Circle $C_A$ has a center at the origin, and radius $r_A = Vmax$, whereas circle $C_B$ has its center at the destination and a radius $r_B = n * V_{max}$ (Fig. 3i). Within the intersection of these two circles (i.e. within the PpA), the new point ($A_1$) is picked generating possible coordinates for the two dimensions, using a uniform random distribution.

Once the point $A_1$ has been created, the same procedure is followed, with the difference that circle $C_A$ now moves its center to $A_1$, for the calculation of point $A_2$, then at $A_2$ for calculating point $A_3$, and so on (Fig. 3ii). In the end, when all the points have been created, the destination will be reached, while respecting both time and speed limitations.

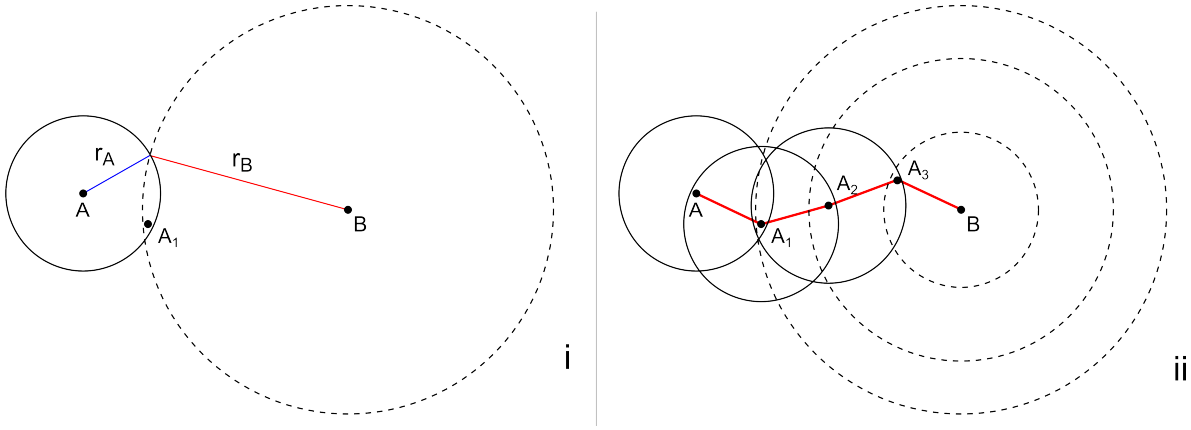



Figure 3. Point generation procedure.

While this algorithm potentially covers many application domains, in ecology the researchers deal with migration patterns that often cover intercontinental movement. For such long-haul distances, the curvature of the earth must be taken into consideration when generating the trajectory. To account for this distortion, the proposed algorithm has been extended by calculating the circle-circle intersection on a spherical surface. The solution is based on spherical trigonometry identities, and uses transposition equations and transformation matrices for performing the necessary calculations.

## 4. Results

The scenario used for the the following experiment was a migrating bird carrying a GPS tag and moving between two distant geographical points. If the GPS data is partially corrupted, or has a significant temporal lag between two consecutive fixes, gaps will need to be filled to ensure the dataset is homogenous—thus comparable—between individuals or species.

Figure 4 depicts the result of generating a single trajectory (left) and 50 random trajectories (right), respectively, from origin *A* to destination *B*, given a speed of 4.9 km/h and an available time of 26 days, 1 h and 10'. Each trajectory consists of 94 points (3-4 points per day). Points *A* and *B* are spaced 800 km apart.

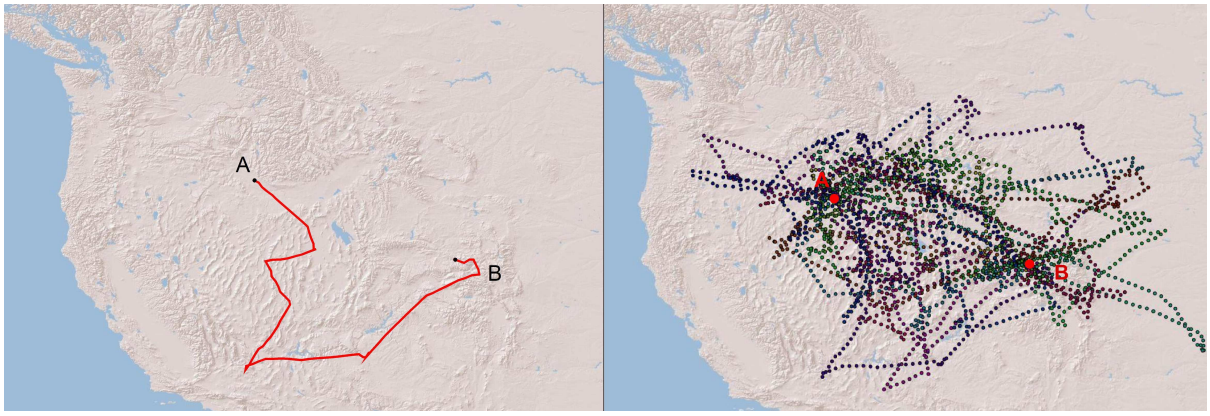



Figure 4. Trajectories generated by the proposed algorithm between origin *A* and destination point *B*, spaced 800 km apart. Left: Single trajectory. Right: 50 simulated trajectories.

We generated 1200 random trajectories between the same points *A* and *B*, which were then rasterized using a line density function with a bandwidth of 0.5 degrees (Fig. 5). It becomes apparent that even with the relatively small number of 1200 trajectories, the density surface starts approaching the space-use surface of the BB movement model. The origin and destination points possess the maximum probability, as expected, since all the simulated trajectories start and end in the same points. Also, the pixels falling near the straight line connecting the endpoints have higher probability of being selected, exactly as in the case of the BB. Similarly, the further a pixel is from this line, the more its probability fades out. Finally, our approach results in a growing region with the two endpoints as centers, and the tendency to approximate an ellipsis, exactly as in the case of BB.

Future research will on the one hand focus on the extended quantitative and qualitative evaluation of the algorithm, and on the other hand, we plan to exploit its modular design and expand it further to account for change of internal state, behavior and context boundaries.

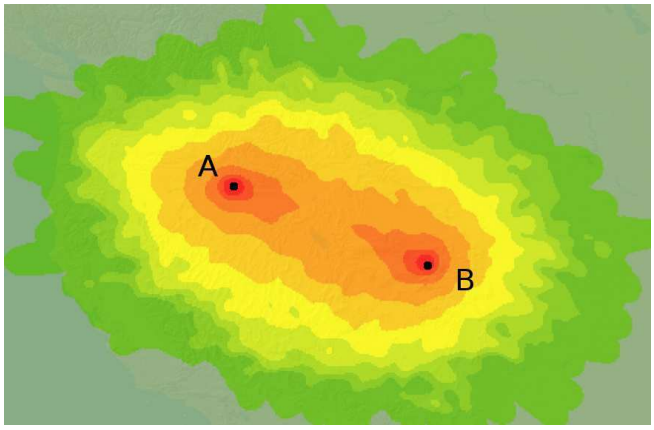



Figure 5. The line density result of 1200 simulated trajectories.

## Acknowledgements

This research represents part of the PhD project of the first author. Funding by the Swiss State Secretariat for Education, Research and Innovation (SERI) through project CASIMO (C09.0167) is gratefully acknowledged. We would like to thank Walied Othman and Kamran Safi for their valuable inputs to this work.

## References

Benhamou S, 2011, Dynamic Approach to Space and Habitat Use Based on Biased Random Bridges. *PLoS ONE*, 6(1): e14592.

Bullard F, 1991, Estimating the Home Range of an Animal: A Brownian Bridge Approach. PhD Thesis. University of North Carolina at Chapel Hill.Bartumeus F, da Luz MGE, Viswanathan, GM, and Catalan J, 2005, Animal search strategies: a quantitative random-walk analysis. *Ecology*, 86(11), 3078–3087.

Codling EA, Plank, MJ, & Benhamou S, 2008, Random walk models in biology. *Interface*, 5(25), 813–34.

Codling EA, Bearon RN, and Thorn GJ, 2010, Diffusion about the mean drift location in a biased random walk. *Ecology*, 91(10), 3106–13.

Fronhofer, EA, Hovestadt, T, & Poethke, H.-J., 2013, From random walks to informed movement. *Oikos*, 11(6), 857–866.

Hägerstrand T, 1970, What about people in regional science? *Regional Science Association. Papers and Proceedings*, 24: 7-21.

Horne JS, Garton EO, Krone SM and Lewis JS, 2007, Analyzing animal movements using Brownian bridges. *Ecology*, 88: 2354-2363.

Hornsby K and Egenhofer M, 2002, Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1–2): 177–194.

Kuijpers B, Grimson R, Othman W, 2011, An analytic solution to the alibi query in the space-time prisms model for moving object data. *International Journal of Geographical Information Science*, 25(2): 293-322.

Miller, HJ, 1991, Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems*, 5(3), 287–301.

Song Y and Miller HJ, 2014, Simulating visit probability distributions within planar space-time prisms. *International Journal of Geographical Information Science*, 28(1): 104–125.

Turchin P, 1998, *Quantitative Analysis of Movement*. Sinauer Associates, Sunderland, MA.

# Investigating the Effects of Activity Space on the Measurement of Segregation using FEATHERS Simulation Data

S.-Y. Hong[1], Y. Sadahiro[1], S.-J. Cho[2]

[1]Center for Spatial Information Science, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba, 277-8568, Japan
Email: {yun.hong; sada}@csis.u-tokyo.ac.jp

[2]Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5 bus 6, 3590 Diepenbeek, Belgium
Email: sungjin.cho@uhasselt.be

## 1. Introduction

Segregation is a reflection of not only individual preference but also the degree of discrimination and socio-economic inequality in our society. Since an accurate assessment of this multidimensional phenomenon is essential for understanding various social problems and conflicts, the measurement of segregation has been an important research topic in the field of geography and sociology.

The recent advances in computing power and the increasing availability of detailed data on daily travel patterns have, in particular, encouraged the development of various individual-level measures of segregation. For example, Wong and Shaw (2011) incorporated the concept of activity space into the exposure index, so it takes into account individual experiences beyond their neighbourhoods. In a similar vein, Farber, Páez, and Morency (2012) demonstrated how the $G_i^*$ statistic and what they termed "relative accessibility deprivation indices" can be used as an individual-level measure of clustering and exposure. While these methods might be able to represent the level of social interaction between population groups more precisely, it is still not clear what their merits and limitations are compared to more traditional, census tract-based indices.

In this extended abstract, we attempt to evaluate the effects of incorporating individual activity spaces into the measure of exposure and isolation. To this end, we employ the activity space-based exposure index developed by Wong and Shaw (2011), but with an adjustment so that the chance of contact between two individuals at a given location is subject to the amount of time they spent together there. This modified index will be applied to a simulated daily travel data set of Seoul, South Korea, and the results will be compared with those from existing methods.

## 2. Methods

### 2.1 Tract-based and Surface-based Exposure Indices

The exposure index estimates the probability of encountering different population groups in a specific geographic area, typically a residential neighbourhood. Earlier studies focused on the level of demographic diversity within residential areas, because it is where most people spend the majority of their time and experience a variety of social interactions.

Census tract-based indices assume that one's neighbourhood is limited to the census areal unit in which the person resides, and that the spatial extent of neighbourhood is identical for all individuals living in the same unit, no matter whether they live close to the edge of the unit, or around the centre of it. More recently developed surface-based measures avoid this

rather unrealistic assumption, as they allow constructing a separate neighbourhood at each point of measurement. In Section 3, we will calculate two indices for income groups in Seoul, the traditional exposure index, $P^*$ (Lieberson 1981), and the spatial exposure index, $\tilde{P}^*$ (Reardon and O'Sullivan 2004), to make comparisons with the activity space-based approach.

The traditional exposure index, $P^*$, is defined as:

$$P^*_{A\times B} = \sum_{i=1}^{n} \left( \frac{A_i}{A} \times \frac{B_i}{T_i} \right) \quad (1)$$

where n is the number of census tracts, A is the total number of minorities in the study region, and $A_i$, $B_i$, and $T_i$ are the population counts of minorities, majorities, and all groups in the unit i.

The spatial exposure index, $\tilde{P}^*$, can be computed in a similar manner:

$$\tilde{P}^*_{A\times B} = \sum_{i=1}^{n} \left( \frac{A_i}{A} \times \tilde{\pi}_{i,B} \right) \quad (2)$$

where n, A, and $A_i$ are the same as in (1), and $\tilde{\pi}_{i,B}$ represents the proportion of the majority population in the neighbourhood of people located at i.

### 2.2 Activity Space-based Exposure Index

Suppose that the activity space of an individual i consists of discrete points that represent the exact locations of daily activities. At each $j^{th}$ activity space, we can calculate the likelihood that an individual of group A contacts members of group B and sum them up with weights proportional to the amount of time the person spent at that location:

$$P^*_{i,A\times B} = \sum_{j=1}^{n} \left( \frac{w_{ij}}{W_i} \times \frac{B_{ij}}{T_{ij}} \right) \quad (3)$$

where n is the number of places i visited, $w_{ij}$ is the amount of time the person i stayed at the $j^{th}$ location, and $W_i$ is the total time recorded for i in the data set (i.e., $W_i = \sum w_{ij}$). $T_{ij}$ is the total population of all groups and is defined as:

$$T_{ij} = \frac{1}{w_{ij}} \sum_{k=1}^{m} \tau_{ijk} \quad (4)$$

where $\tau_{ijk}$ refers to the amount of time i and k spent together at j (k = 1, 2 … m). Note that (3) is similar to the individual-level analogue of the exposure index presented in Wong and Shaw (2011), but only those who stay at j during the same time frame as i are accounted for the calculation of this measure.

## 3. Segregation by Income in Seoul

### 3.1 Data

This study employs an individual trajectory data set that describes daily routines of 21,266 households and 53,002 individuals in the Seoul Metropolitan Area (SMA), South Korea. The daily routine data were obtained from the FEATHERS simulation (Bellemans et al. 2010), which predicts individual schedule based on the relationship between people's socio-economic characteristic and activity-trip behaviour. The data produced from this simulation program encompass various socio-economic attributes for households and their members (e.g., household income, the number of private vehicles in the household, age, gender, and work status of each member, etc.), as well as information on their daily routines. In this

extended abstract, we use the household income data to classify the individuals into two groups, low-income group (i.e., < 2,000,000 Korean Won per month) and middle- and upper-income groups (Figure 1), and examines the exposure level between them.

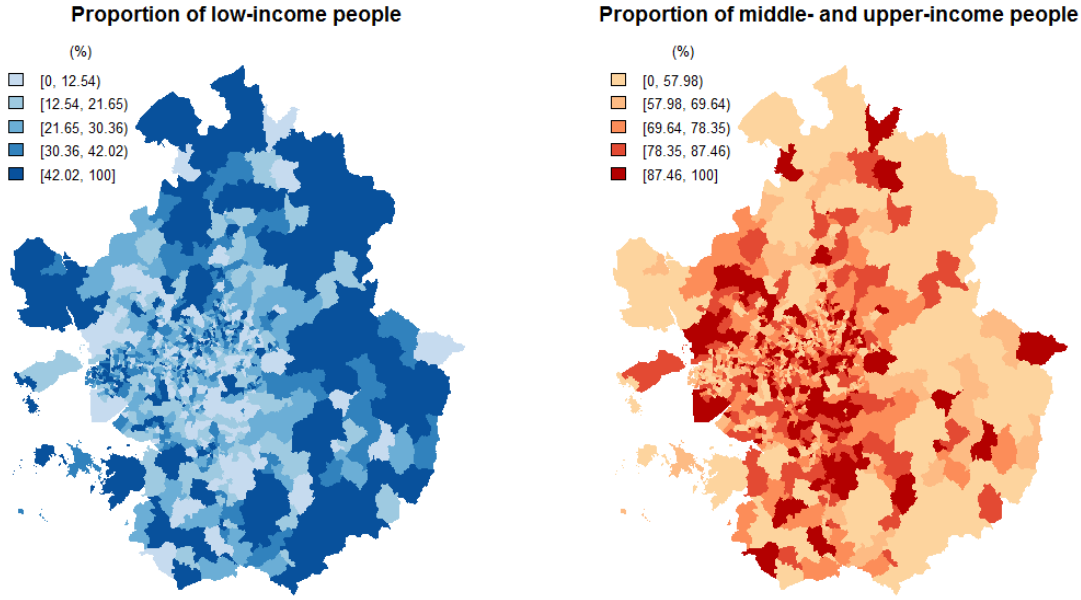



Figure 1. Proportions of low-income people (left) and middle- and upper-income people (right) in the SMA at the census tract level.

### 3.2 Results

The level of exposure between low-income group (A) and middle- and upper-income groups (B) was measured by three different indices, $P^*$, $\widetilde{P}^*$, and the activity space-based exposure index, $AP^*$. It should be noted that the first two measures evaluate the degree of residential segregation. As shown in Table 1, $P^*$ and $\widetilde{P}^*$ produced similar figures, but $AP^*$ was considerably higher for both A to B and B to A.

This result is probably because most of the people travelled to the central business district and stayed there during the daytime for work. It would have made a space where people with diverse economic backgrounds are mixed, consequently moderating the degree of segregation (i.e., increases the level of exposure to different income groups). Figure 2 displays that the level of exposure has increased in almost all census tracts compared to Figure 1.

Table 1. The level of exposure between low-income group (A) and middle- and upper-income groups (B), measured by three different indices: $P^*$, $\widetilde{P}^*$, and $AP^*$.

| Method | A to B | B to A |
|---|---|---|
| $P^*$ | 0.6583 | 0.2171 |
| $\widetilde{P}^*$ | 0.6654 | 0.2203 |
| $AP^*$ | 0.7525 | 0.2475 |

## 4. Discussion

The widespread of GPS-enabled devices and other data collection techniques over the past few decades has led to the availability of micro-level demographic data, and it has permitted

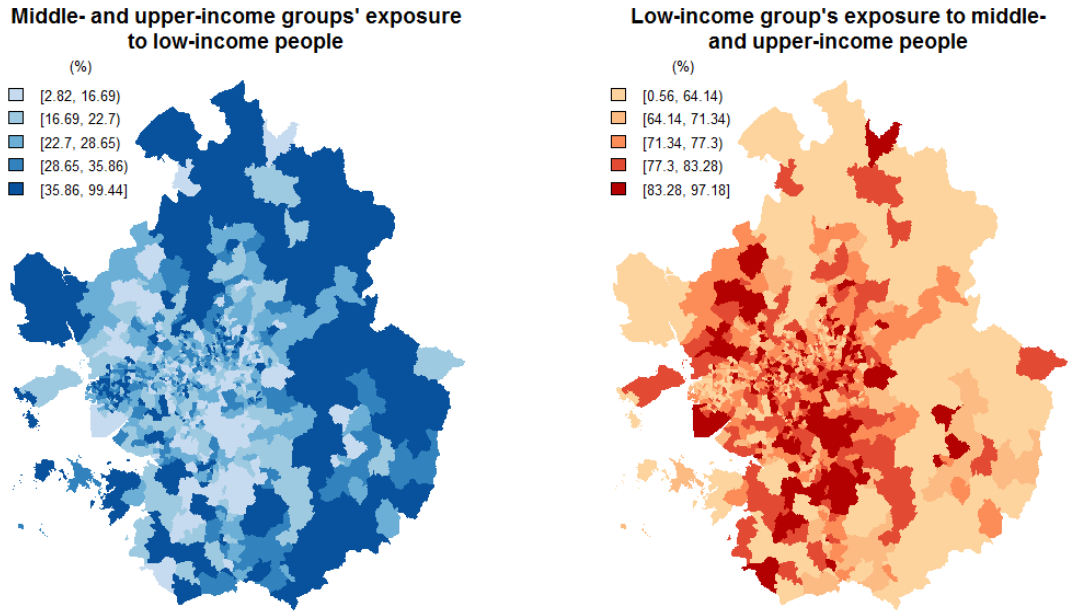



Figure 2. Varying exposure of middle- and upper-income groups to low-income group (left) and the reverse (right) across census tracts.

the development of segregation measures that explicitly consider the spatial extent of individuals' daily activity patterns. Unlike the conventional, tract-based methods, such activity space-based indices do not rely on census areal units as the context of their everyday lives, so this might be able to capture the reality of people's experiences more accurately.

However, this sort of approach should be used with care, because one's activity space consists of many different types of places. Depending on where we are—at home, work, school, café, church, or anywhere in between—we interact with people around us differently. Ideally, for each place, the degree of exposure/isolation should be measured in a different way that reflects the purpose of the visit and the characteristics of the place. However, most, if not all, of the current implementations do not differentiate between places, and this could result in under- or overestimation of the people's actual experiences. As demonstrated in the previous section, the amount of time spent at each location can be used as an indirect indicator of that place's significance, but further research would be needed to develop a more reliable activity space-based exposure index.

## Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24-02309.

## References

Bellemans T, Kochan B, Janssens D, Wets G, Arentze T and Timmermans H, 2010, Implementation framework and development trajectory of FEATHERS activity-based simulation platform. Transportation Research Record: Journal of the Transportation Research Board, 2175(1):111–119.

Farber S, Páez A and Morency C, 2012, Activity spaces and the measurement of clustering and exposure: a case study of linguistic groups in Montreal. Environment and Planning A, 44(2):315–332.

Lieberson S, 1981, An Asymmetrical Approach to Segregation. In: Peach C, Robinson V and Smith S (eds), Ethnic Segregation in Cities. Croom Helm, London, UK.

Reardon S and O'Sullivan D, 2004, Measures of spatial segregation. Sociological Methodology, 34(1): 121–162.

Wong DWS and Shaw SL, 2011, Measuring segregation: an activity space approach. Journal of Geographical Systems, 13(2):127–145.

# Modularity and spectral regional clustering by commuter flows

C. Kaiser, M. Kordi, F. Bavaud

Institute of Geography and Sustainability, University of Lausanne, Switzerland
Email: {christian.kaiser , maryam.kordi , francois.bavaud}@unil.ch

## 1 Introduction

Strongly mutually spatially interacting regions form natural groups of regions, supposedly well-connected in view of the distance-decaying nature of spatial interaction. The latter could be primarily measured by flow between pairs of regions, such as commuter flows. Turning those considerations into operational clustering algorithms necessitates to confront two related difficulties, namely the definition of spatial interaction from flows, and the definition of the clustering objective.

This contribution considers modularity, $K$-means and spectral clustering with the example of Swiss communes derived from commuter flows, as defining the symmetric network specified by the spatial weights provided by the normalised *exchange matrix*. The latter can in addition be renormalised by over- or under-weighting the contribution of large populated regions.

While we consider commuter flows in this paper, the presented considerations apply also to other types of spatial interaction flows, for example flows of goods or migration.

## 2 Formalism: definitions and analysis

Commuter flow $T$ arguably generates a weighted network, whose *affinity matrix*, fixing the edge values, can as a first estimate be defined as proportional to the corresponding daily commuters count - a generally *asymmetric* prescription. By contrast, this contribution considers *symmetric* and normalised affinity matrices, referred to as *exchange matrices* $E = (e_{ij})$ (Berger and Snell, 1957), and obeying

$$e_{ij} = e_{ji} \geq 0 \qquad e_{i\bullet} = \sum_{j=1}^{n} e_{ij} = f_i \qquad e_{\bullet\bullet} = f_\bullet = \sum_{j=1}^{n} f_i = 1 \ . \tag{1}$$

Here and in the sequel, ``$\bullet$" denotes the summation over the replaced index, as in $e_{i\bullet} = \sum_{j=1}^{n} e_{ij}$. In this setup, $e_{ij}$ is the relative weight of the pair $ij$, and $f_i$ the relative weight of region $i$; they respectively constitute pair and individual probabilities.

Clustering symmetrical networks has been plentifully investigated within two paradigms, namely $K$*-means* and *spectral clustering* on one hand, and *modularity maximisation* on the other hand (see e.g. Schaeffer, 2007; White and Smyth, 2005; Malliaros and Vazirgiannis, 2013; Fortunato, 2010). Their mutual relationship, less explored, is

compared below in the *exchange matrix* setup, compatible with weighted regions and fuzzy membership - a fairly general formalism, made natural by aggregation-invariance considerations.

## 2.1 Membership matrix $Z$. Modularity and $K$-means $E$-clustering

A general fuzzy or *soft* clustering with $m$ groups is given by the $n \times m$ non-negative *membership matrix* $Z = (z_{ig})$, with $z_{i\bullet} = 1$, interpretable as the probability that region $i$ belongs to group $g$. Membership matrices yield group *weights* $\rho_g$, group *overlap* $\theta_{gh}$ and group *associations* $a_{gh}$ as

$$\rho_g = \sum_{i=1}^{n} f_i z_{ig} \qquad \theta_{gh} = \sum_i f_i z_{ig} z_{ih} \qquad a_{gh} = \sum_{ij} e_{ij} z_{ig} z_{jh} \ . \tag{2}$$

By construction, $\theta_{gg} \leq \rho_g$, with equality iff the clustering is hard, that is iff $z_{ig}^2 = z_{ig}$. The *cut* of group $g$ is the total association of $g$ with the *other* groups, interpretable as measure of *surface* of $g$, while $\rho_g$ is interpretable as a measure of its *volume*:

$$\texttt{CUT}_g := \sum_{h \mid h \neq g} a_{gh} = \sum_{ij} e_{ij} z_{ig} (1 - z_{jg}) = \rho_g - a_{gg} \qquad \texttt{VOL}_g := \rho_g \ . \tag{3}$$

The *normalised cut* or NCUT associated to a network $E$ partitioned as $Z$ measures the relative overlap between the different groups, or, oppositely, the average group self-isolation or *normalised association* NASS (Shi and Malik, 2000):

$$\texttt{NCUT}[Z] = \sum_g \frac{\texttt{CUT}_g}{\texttt{VOL}_g} = \sum_{g=1}^{m} \frac{\rho_g - a_{gg}}{\rho_g} = m - \texttt{NASS}[Z] \qquad \texttt{NASS}[Z] := \sum_{g=1}^{m} \frac{a_{gg}}{\rho_g} \ . \tag{4}$$

On the other hand, the *modularity* $Q$, introduced by Newman (2004) for hard partitions, is expressed in the present soft exchange-based setup as

$$Q[Z] = \sum_{ij} (e_{ij} - f_i f_j) \sum_g z_{ig} z_{jg} = \sum_{g=1}^{m} (a_{gg} - \rho_g^2) \ . \tag{5}$$

Both $\texttt{NASS}[Z]$ and $Q[Z]$ constitute distinct measures of the quality of the network partitioning, taking large values provided $Z$ is a good clustering, i.e. $Z$ manages grouping pairs of nodes $ij$ possessing large mutual values $e_{ij}$.

A third $K$*-means clustering criterion* should be added, provided that $E$ is *positive semi-definite*, that is possessing only non-negative eigenvalues: in that case, the *diffusive dissimilarity* (Bavaud, 2014)

$$\mathcal{D}_{ij}[E] := \frac{e_{ii}}{f_i^2} + \frac{e_{jj}}{f_j^2} - 2\,\frac{e_{ij}}{f_i f_j} \tag{6}$$

behaves as a squared Euclidean dissimilarity between regions. The corresponding *within-groups inertia* can, with simple algebra (e.g. Bavaud and Cocco, 2013), be expressed as

$$\epsilon_{ij}[Z] := f_i f_j \sum_g \frac{z_{ig} z_{jg}}{\rho_g} \qquad \Delta_W[Z,E] := \frac{1}{2} \sum_{ij} \epsilon_{ij}[Z]\, \mathcal{D}_{ij}[E] = \sum_i \frac{e_{ii}}{f_i} - \texttt{NASS}[Z,E] \tag{7}$$

that is, minimising $\Delta_W[Z]$ amounts in maximising $\texttt{NASS}[Z]$ (Dhillon et al., 2004). Hence, $K$-means and spectral clustering criteria are identical, yet still distinct of the modularity criteria at this stage.

### 2.2 Commuters exchange $E(T)$ and renormalisation

Determining an exchange matrix $E$ from flows $T$ is a central question, whose answer has elicited various proposals. The simple choice

$$E = \tilde{T}\Pi^{-1}\tilde{T} \qquad \text{where} \qquad \tilde{T} = \frac{T + T'}{\mathbf{1}'(T + T')\mathbf{1}} \qquad \text{and} \qquad \Pi = \text{diag}(\tilde{T}\mathbf{1}) = \text{diag}(f) \tag{8}$$

makes $E$ non-negative, symmetric, normalized, and *positive semi-definite* (p.s.d.). Above $\mathbf{1}$ is the unit vector and $T'$ is the transpose of $T$.

Once $E$ is computed, modularity maximisation can be implemented through the function `label.propagation.community()` of the `R` package `igraph`, taking as argument the exchange matrix $E$ itself, which defines the edge weights of the unoriented graph (Csardi and Nepusz, 2006).

$K$-means clustering is implemented through the function `kkmeans()` of the `R` package `kernlab` (Karatzoglou et al., 2004), taking as argument the ``linear kernel" $S$ consisting of the non-centred scalar product associated to the diffusive distance (6), namely $S = \Pi^{-1}E\Pi^{-1}$.

A generalised family of $K$-means clusterings results form the consideration of *renormalised kernels* defined as $S(\kappa) = \Pi^{-\kappa}E\Pi^{-\kappa}$, amounting in considering *renormalised diffusive dissimilarities* and *renormalised associations*

$$\mathcal{D}_{ij}[E,\kappa] = \frac{e_{ii}}{f_i^{2\kappa}} + \frac{e_{jj}}{f_j^{2\kappa}} - 2\,\frac{e_{ij}}{(f_i f_j)^{\kappa}} \qquad \texttt{NASS}[Z,\kappa] := \sum_{g=1}^{m} \frac{a_{gg}(\kappa)}{\rho_g} \qquad a_{gg}(\kappa) = \sum_{ij} \frac{e_{ij} z_{ig} z_{jg}}{(f_i f_j)^{\kappa-1}} \tag{9}$$

Unrenormalised $K$-means obtains for $\kappa = 1$. The smaller $\kappa$, the larger the contribution from less populated regions with $f_i << 1$. The case $\kappa = 0.5$ is of special interest, since $e_{ij}/\sqrt{f_i f_j}$ constitutes, in the limit $f_i, f_j << 1$, the decision variable for testing the independence of the corresponding flow $H_0 : e_{ij}^{\text{theo}} = f_i^{\text{theo}} f_j^{\text{theo}}$ (see e.g. Haberman, 1973); also, the renormalised within-inertia with $\kappa = 0.5$ in (9) arguably resembles the modularity criterion (5) (when unduly neglecting a factor $\rho_g/\sqrt{f_i f_j}$ multiplying the exchange matrix).

## 3 Case study

To illustrate the presented considerations, we use as an example the journey-to-work flows for Switzerland for the year 2000. The dataset contains roughly 187,000 non-zero flows between all 2896 Swiss communes, a commune being the smallest administrative division in the country. After elimination of intra-zonal flows (workers living and working in the same commune), about 184,000 non-zero flows remain in the dataset.

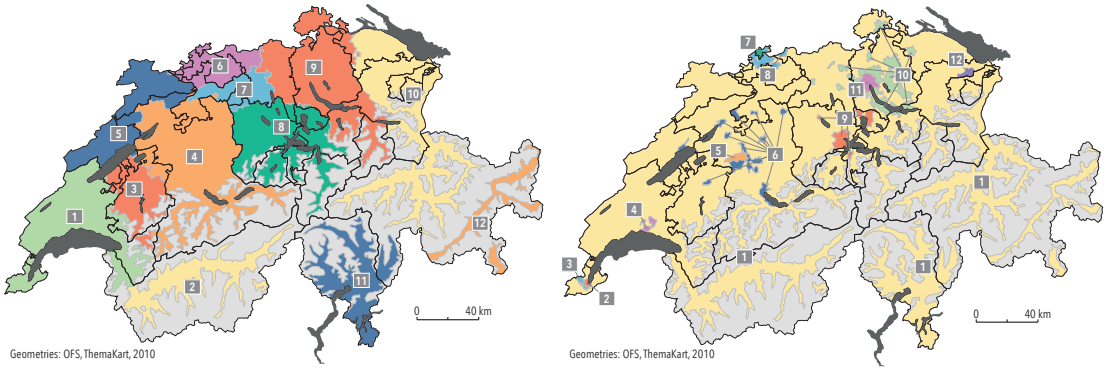



Fig. 1: Left: modularity clustering (``fast-greedy" variant; see Csardi and Nepusz 2006), taking on its maximum for $m = 12$ groups. Right: renormalised kernel $K$-means for $\kappa = 0$ with $m = 12$ groups.

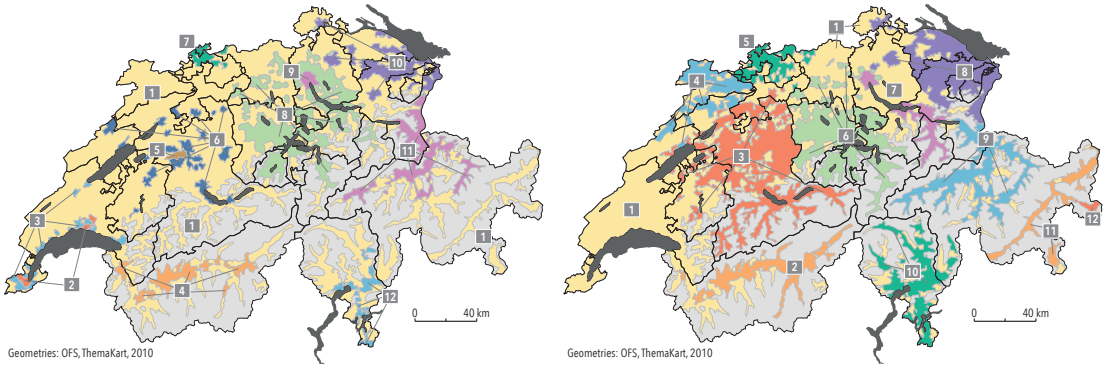



Fig. 2: Renormalised kernel $K$-means for $\kappa = 0.25$ (left) and $\kappa = 0.5$ (right). Groups are still disconnected, but their spatial extension appears more uniform.

For each flow, the number of commuters is known. The dataset is provided by the Swiss Federal Statistical Office (SFSO) within the framework of the population census and is freely available on `www.pendlerstatistik.admin.ch`. In order to obtain comparable clusters, the number of groups has been set to $m = 12$ for all clustering variants, which is the number of clusters for the best partitioning using the modularity approach.

The maps in figures 1 to 3 show the results of the two different clustering approaches and a varying $\kappa$ value for the $K$-means approach. On all maps, mountainous areas without population are represented in light gray. The outlines of the 26 cantons, the most widely used administrative division in the country, are shown as thin black lines. Clusters are shown in arbitrary colours and are given a random number between 1 and 12.

Modularity clustering (figure 1 left) appears to produce well-connected groups of balanced size. These regions are similar to some extent to the statistically defined labour market regions (Schuler et al., 2005). By contrast, renormalised kernel $K$-means for

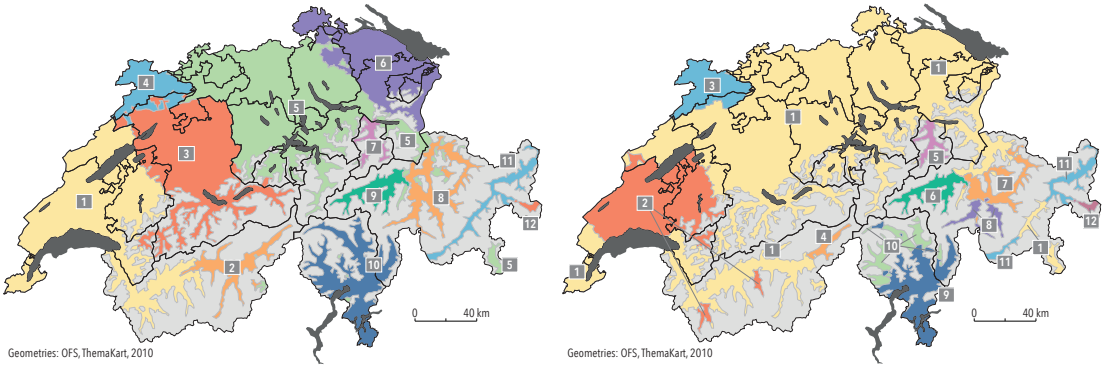



Fig. 3: Renormalised kernel $K$-means for $\kappa = 0.75$ (left), arguably comparable to Figure 1 left, and $\kappa = 1$ (right) shows mainly remote closed labour markets.

$\kappa = 0$ (figure 1 right) strongly overweights the contribution of communes with large weights $f$. As a consequence, the biggest cities (i.e. Zürich, Bern, Basel, Lausanne and Geneva) form a cluster on their own, and smaller cities around these big agglomerations are grouped together in additional clusters, leaving the lesser populated territory in a single giant cluster (2797 out of the 2896 communes with a population of 4,7 million out of 7,3 million). By increasing the $\kappa$ parameter, this effect can be reduced. For a value of $\kappa = 0.75$ (figure 3 left), the groups are getting comparable to those found with the modularity clustering (figure 1 left). For values $\kappa = 0.25$ and $\kappa = 0.5$ (figure 2), the groups are showing mostly the dominance of the big cities, but the regional labour market patterns are emerging. In the case of $\kappa = 1$, mainly remote alpine valleys are making up the biggest number of clusters. Due to the topography, transportation in these valleys is not very fast, and journey-to-work trips to other valleys are unlikely. As a result, nearly closed local labour markets emerge, which are detected with a $\kappa$ value of 1.

## 4 Conclusion

This paper has shown links and differences between several well-known clustering methods applied to the problem known in network theory as community detection. An important aspect is how to build a non-negative, symmetric, normalised exchange matrix $E$ from the flows. Spectral clustering and k-means clustering criteria are identical, while the modularity criterion is similar up to a factor only after renormalisation.

A case study for the journey-to-work flows in Switzerland illustrates the considerations on the relationship between k-means and modularity-based clustering. The regionalisation based on commuter flows yields in some circumstances clusters similar to labour market regions. However, it has to be noted that clustering is only one possibility to define labour market regions and that in some cases other methods are used or clustering is only used for some parts. See for example Schuler et al. (2005), Eckey et al.

(2006) or Kropp and Schwengler (2011) for considerations around defining labour market regions for Switzerland respectively for Germany. However, it is interesting to note that modularity clustering and for appropriate values of $\kappa$ also k-means clustering yield spatially connected clusters even without corresponding constraints. Also, for different values of $\kappa$, different structures underlying the journey-to-work flows become visible, such as local labour markets.

Finally, we would like to repeat that the presented considerations do not only apply to commuting, but to any spatial interaction flow.

## References


Bavaud, F. (2014). Spatial weights: constructing weight-compatible exchange matrices from proximity matrices. In *Proceedings of the GIScience 2014 conference, Vienna*.

Bavaud, F. and Cocco, C. (2013). Factor analysis of local formalism. In *Proceedings of the European Conference on Data Analysis 2013, Luxembourg*.

Berger, J. and Snell, J. L. (1957). On the concept of equal exchange. *Behavioral Science*, 2(2):111--118.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, page 1695.

Dhillon, I., Guan, Y., and Kulis, B. (2004). A unified view of kernel k-means, spectral clustering and graph cuts. Technical Report TR-04-25, Computer Science Department, University of Texas at Austin.

Eckey, H.-F., Kosfeld, R., and Türck, M. (2006). Abgrenzung deutscher Arbeitsmarktregionen. *Raumforschung und Raumordnung*, 64(4):299--3009.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75--174.

Haberman, S. (1973). Log-linear models for frequency data: Sufficient statistics and likelihood equations. *The Annals of Statistics*, 1(4):617--632.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab -- an S4 package for kernel methods in R. *Journal of Statistical Software*, 11:1--20.

Kropp, P. and Schwengler, B. (2011). Abrenzung von Arbeitsmarktregionen - ein Methodenvorschlag. *Raumforschung und Raumordnung*, 69(1):45--62.

Malliaros, F. D. and Vazirgiannis, M. (2013). Clustering and community detection in directed networks: a survey. *Physics Reports*, 533(4):95--142.

Newman, M. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27--64.

Schuler, M., Dessemontet, P., and Joye, D. (2005). *Die Raumgliederungen der Schweiz*. Bundesamt für Statistik (BFS), Neuchâtel.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888--905.

White, S. and Smyth, P. (2005). A spectral clustering approach to finding communities in graphs. In Kargupta, H., Srivastava, J., Kamath, C., and Goodman, A., editors, *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 274--285.

# Point Process Models for Prospective Crime Analysis

G. Rosser[1], T. Cheng[1]

[1] SpaceTimeLab, Department of Civil, Environmental & Geomatic Engineering, University College London, London WC1E 6BT
Email: {g.rosser; tao.cheng}@ucl.ac.uk

## 1. Introduction

The criminological theory of near repeat victimisation states that the occurrence of certain crimes increases the risk of further crimes within the local neighbourhood for some ensuing time period (Johnson and Bowers, 2004; Youstin et al., 2011). This leads to a temporally and spatially localised cluster of crime events following an initial background event. The real-time identification of emerging crime 'hotspots' has been the target of much research effort (S. Chainey et al., 2002; Bowers et al., 2004), as such a tool would be of great utility to police forces worldwide. The majority of existing methods use statistical analysis of crime data to infer the presence of hotspots. Such an approach is valuable, however it is not generally suitable for generating future predictions as the methods are not based on well-stated models, instead relying on heuristics. Furthermore, such analyses give little insight into the underlying method of generation of crime patterns.

In this report, we consider the application of a self-exciting point process (SEPP) model to crime data. The point process framework is a good description for crime data, which comprise a series of geolocated points in time. Methods based on point processes have previously been developed to detect space-time clustering (Diggle et al., 1995), however such statistical approaches are more suited to retrospective analysis than forecasting. In a promising recent development in the field of criminology, Mohler et al. adapted a point process method for predictive modelling of crime data (Mohler et al., 2011). The authors use a modification to a well-established SEPP model describing the spread of seismic activity known as Epidemic Type Aftershock Sequences (ETAS). Their method outperforms a kernel-based hotspot detection approach in terms of predictions made on real crime data.

Despite the success of the approach taken by Mohler et al. in analysing and predicting crime, there are several open questions and issues preventing the widespread adoption of the method. First, the approach is computationally intensive due to the necessity of evaluating a kernel density estimation (KDE) at a large number of points (typically millions to tens of millions per iteration). Second, the KDE used by Mohler et al. assumes a multivariate Gaussian kernel, which is invalid in the temporal dimension as it permits crimes to be triggered by future events. Finally, there is no open source implementation of the SEPP, which hampers further research and development of the methods discussed.

The subject of this abstract is the development and implementation of a robust computational tool to apply the SEPP to crime data. We assess the predictive performance of our method using appropriate validation methods, such as the measure of search efficiency rate. We apply our method to open crime data provided by the city of Chicago, USA, to demonstrate its effectiveness. The analysis of a non-public domain crime dataset provided by the Metropolitan Police Service (MPS), London, UK is a current work in progress.

# 2. Materials and Methods

## 2.1 Self-exciting point process

At the core of the SEPP model of crime is the conditional intensity, $\lambda(t,x,y)$, which gives the density of the expected rate of occurrence of crimes in a small neighbourhood around the region $(x,y)$ at time $t$, conditional upon the history of all occurrences up to that time. The conditional intensity may be described as the sum of background and triggered events:

$$\lambda(t,x,y) = \mu(t,x,y) + \sum_{\{k:t_k<t\}} g(t-t_k, x-x_k, y-y_k), \tag{1}$$

where $\mu$ denotes the background occurrence rate and $g$ denotes the triggering kernel. Thus all crimes that have occurred prior to a given time may theoretically contribute some additional expectation of the current crime activity, though in practice this may vanish over some period of time and/or distance.

In order to apply this theory to real data we must estimate the functional forms of $\mu$ and $g$. This problem reduces to one of declustering the data (Zhuang et al., 2002) in order to identify which events in a crime dataset arise from the background activity and which have been triggered by previous events. Various approaches have been proposed in the seismology literature, commonly involving maximum likelihood approaches based on assumptions of the forms of $\mu$ and $g$ (Daley and Vere-Jones, 2003). A recently developed alternative uses a nonparametric kernel density estimate (KDE) to avoid this necessity (Zhuang, 2006). Mohler et al. adopt this approach in their study (Mohler et al., 2011).

Let $p_{ji}$ denote the probability that event $i$ was triggered by event $j$. By convention, $p_{ii}$ denotes the probability that event $i$ is a background event. Furthermore, $p_{ji} = 0$ if $t_i < t_j$, so that all of the probabilities may be encoded in an upper triangular matrix $P$. Under the assumptions of equation (1), these probabilities are given by

$$p_{ii} = \frac{\mu(t_i,x_i,y_i)}{\lambda(t_i,x_i,y_i)} \tag{2}$$

$$p_{ji} = \frac{g(t_i-t_j, x_i-x_j, y_i-y_j)}{\lambda(t_i,x_i,y_i)}. \tag{3}$$

In (Mohler et al., 2011), an optimisation routine is proposed in which the background and parent/child events are sampled randomly from the data using the probabilities in $P$. Based on these sampled populations, a KDE is computed and $P$ is updated following equations (2) and (3). This algorithm has been shown using simulated data to converge to correct estimates of $\mu$ and $g$.

## 2.2 City of Chicago crime data

For this study we are using crime data available on the City of Chicago's online data portal at https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. The data are extracted from the Chicago Police Department's analysis and reporting system; each crime entry has an associated date, time and geographic location. Data are available from 2001 to the present.

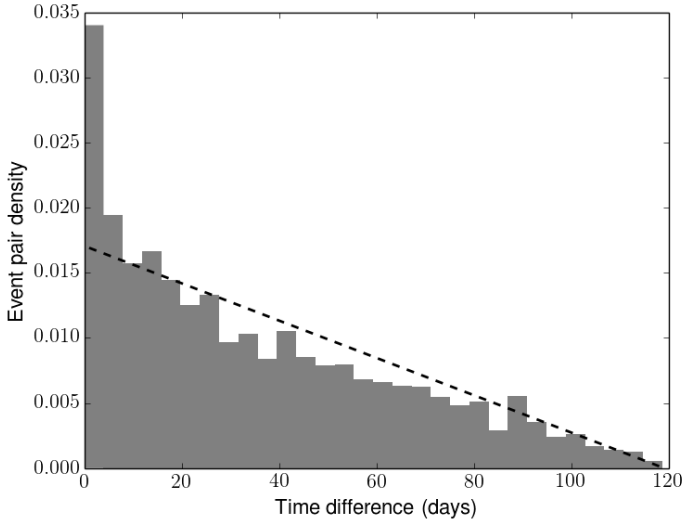



Figure 1: Pairwise time differences for burglaries occurring within 100m of one another in Chicago, first half of 2003. Dashed line indicates the expected distribution if events are produced by a stationary Poisson process.

# 3. Results

## 3.1 Preliminary data analysis

Figure 1 shows the distribution of time differences between pairs of burglaries that occur within 100m of one another in Chicago. Under the assumption that crimes occur as a homogeneous stationary Poisson process, we expect a uniform distribution of crimes in time. This would result in a linearly decreasing distribution of pairwise time differences, as shown. The distribution observed is skewed towards small time differences, suggesting that crimes occur with significant temporal correlation. This departure from linearity diminishes as larger spatial distances are considered.

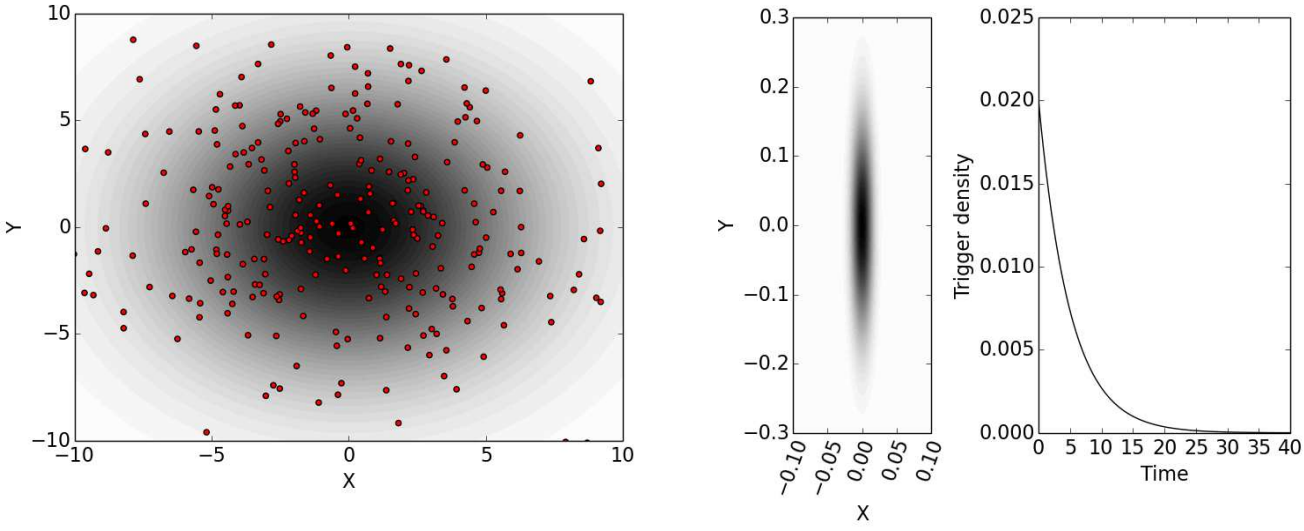



Figure 2: (Left) Background density used in the simulation, overlaid with simulated events (red circles). (Right) Trigger density in space and time.

### 3.2 Simulated data

In order to validate our method, we simulate a crime process with known background and triggering functions (see Figure 2). We use the simulated data to validate the SEPP method, assessing how well it is able to identify background and triggered events.

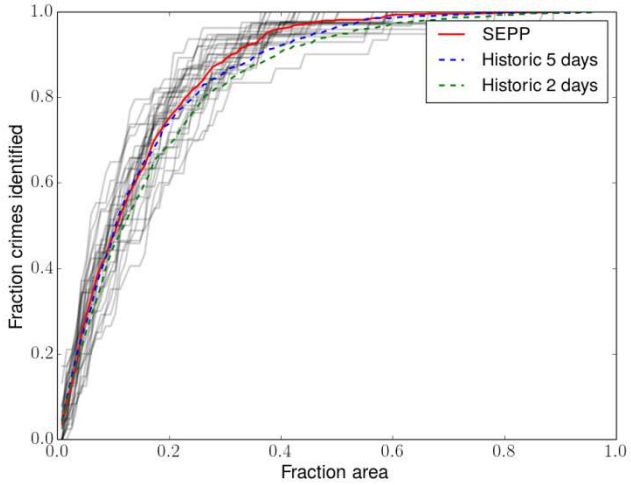



Figure 3: Search efficiency rate plot for simulated data showing the predictive performance of the SEPP compared with the historic heatmap method.

### 3.2 Validation

Figure 3 shows the search efficiency rate of the SEPP applied to the simulated data, defined as the fraction of crime detected as the fraction of area coverage is increased. The SEPP is compared with a historic heatmap approach, in which data from a fixed number of days are used to generate a spatial KDE that is used as a forward prediction. The SEPP performs marginally better than the heatmap approach. Note that the simulated data are generated from a simple model; real crime data are expected to highlight a greater difference between the methods. Table 1 demonstrates that the SEPP infers the lineage of the simulated data with low error rates.

Table 1: Confusion matrix generated by applying the SEPP to simulated data.

| | | **Predicted** | |
|---|---|---|---|
| | | Background | Trigger |
| **Actual** | Background | 1565.2 | 84.7 |
| | Trigger | 29.8 | 377.3 |

### 3.3 Chicago density maps

Density maps for the background component and combined background / trigger components are shown in Figure 4.

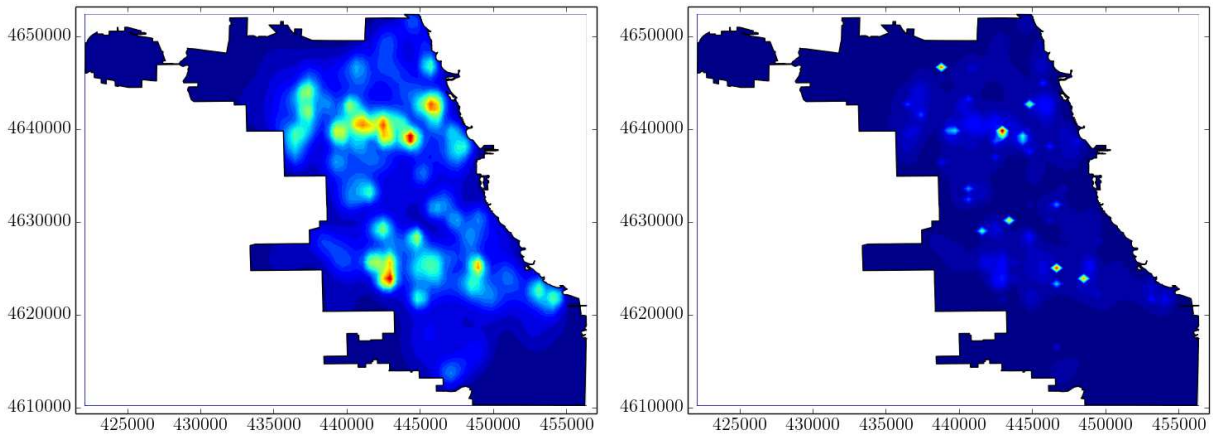



Figure 4: Density maps computed using the SEPP for burglaries in the Chicago region in February 2001. (Left) background density; (right) combined background / trigger density.

## 4. Work in progress

- Full assessment and validation of the SEPP on Chicago data.
- The effect of using different kernels.
- Application of the SEPP to MPS crime data.
- Investigation of which crime types are best described by the SEPP.

## Acknowledgements

This work is part of the project Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks (www.ucl.ac.uk/cpc), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1). The data are generously provided by Metropolitan Police Service (London).

## References

Bowers KJ, Johnson SD and Pease K, 2004. Prospective hot-spotting: The future of crime mapping? British Journal of Criminology, 44:641-658.

Chainey S, Reid S and Stuart N, 2002. When is a hotspot a hotspot? In Socio-Economic Applications of Geographic Information Science, CRC Press.

Daley D and Vere-Jones D, 2003. An Introduction to the Theory of Point Processes (2nd ed.), Springer, New York.

Diggle PJ, Chetwynd AG, Haggkvist R and Morris SE, 1995. Second-order analysis of space-time clustering. Statistical Methods in Medical Research, 4:124.

Johnson SD and Bowers KJ, 2004. The burglary as clue to the future: The beginnings of prospective hot-spotting. European Journal of Criminology, 1:237.

Mohler GO, Short MB, Brantingham PJ, Schoenberg FP and Tita GE, 2011. Self-exciting point process modelling of crime. Journal of the American Statistical Association, 106(493):100-108.

Youstin TJ, Nobles MR, Ward JT and Cook CL, 2011. Assessing the generalizability of the near repeat phenomenon. Criminal Justice and Behaviour, 38:1042.

Zhuang J, Ogata Y and Vere-Jones D, 2002. Stochastic declustering of space-time earthquake occurrences. Journal of the American Statistical Association, 97(458):369-380.

Zhuang J, 2006. Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. Journal of the Royal Statistical Society B, 68(4):635-653.

# Data Imputation in Short-Run Space-Time series – a Bayesian Approach

Chris Brunsdon[1], Martin Charlton[1]

[1] National Centre for Geocomputation, National University of Ireland Maynooth
Email: {christopher.brunsdon;martin.charlton}@nuim.ie

## 1. Introduction

A challenging problem arises when the analyst is presented with incomplete data. An example of this is the population estimates available for different levels in the Nomenclature of Territorial Units for Statistics (NUTS) hierarchy of areal units from EUROSTAT. The NUTS heirarchy consists of a set of several nested sets of spatial divisions from NUTS0 (countries) to NUTS3 (small regions), intended for the purposes of harmonising EU regional statistics, socio-economic analysis and the framing of EU regional policies. The EUROSTAT table *demo_r_pjanaggr3* contains estimates of the population on 1st January from 1990 to 2012 inclusive. Substantial portions of the data are missing particularly for NUTS2 and NUTS3 regions for the earlier part of the time period. The task here is the completion of these data series, by estimating the missing data items.

In some cases comparable data is available from some other data sources, such as the national statistical agency in each country. These values can be compared with the EUROSTAT values where the data *is* available and where there is agreement may be used with some confidence. However, when there is a divergence then the question arises of how best to use the data from the national agency. In spite of availability of additional data, there are still gaps within some of the alternative data series, most notably at the NUTS3 level. The aim of this work is to estimate these quantities.

However, there are a series of issues. First, we are dealing with time series of short duration – no more than 22 years. Second, the series belong to a spatial hierarchy, so any estimates for lower level regions are constrained by estimated or actual values at the higher levels. Third, there is the choice of imputation method. A fourth issue concerns the implementation of the chosen method.

## 2. Time Series Analysis and Missing Data

A time series is a set of ordered measurements of some characteristic taken at regular time intervals. The objectives for time series analysis include analysis, explanation, and prediction (Chatfield 1984). The shortness of the series under consideration here and the pattern of missing data suggests that these traditional approaches to analysis are not suitable – there is insufficient data to provide reliable estimates for the parameters of a classical time series model such as an *Auto Regressive Integrated Moving Average* (ARIMA) model. For example, in a particular NUTS3 region, data might be missing for intercensal years in the 1990s: 1990, and 1992 to 2000 inclusive, so that the evidence we have is the single value in 1991, and the series from 2001 to 2012. However, data may be present for the containing NUTS2 region for the entire time period. The problem becomes one of identifying a technique which will allow us to make use of all the available evidence in a coherent and consistent fashion. A

further unusual contraint is that the summed populations for the NUTS3 regions in a given NUTS2 must equal that of the parent region. This requirement yields further evidence which can be used to ensure the reliability and consistency of the lower level forecasts.

There are also a number of possible approaches to completing the series using simpler algorithms. For example the *hotdeck* and *last observation carried forward* (LOCF) (Enders 2010) approaches. Taking the case of LOCF applied to time series, the data value in the last non-missing time period is copied into the missing parts of the series. However this often fails to encapsulate the underlying missing data generating processes satisfactorily — for example, if there is a population growth trend in the data, LOCF would fail to reflect this, as it would fix missing estimates to be the same as an earlier year.

Thus, a model-based approach is desirable. If we can model the trend in the existing data, and then any autocorrelations in the residuals after the trend is removed, then we have the basis for estimating both any missing data, but also providing an estimate of the uncertainty. As we have noted above, the series are too short for traditional time series methods. For these reasons, we choose a forecasting strategy based on Bayesian methods, and a simpler time series model. In the next section, the Bayesian approach will be briefly outlined.

### 2.1 Bayesian Inference

If we have a set of data $D$ and a set of model parameters $\theta$, then the data is modelled by the probability distribution

$$P(D|\theta)\ . \tag{1}$$

Here both $D$ and $\theta$ may be vectors or matrices as well as scalars. For example $\theta$ could be the triplet of slope, intercept and error variance in a bivariate regression model; in this case $D$ would be a $2 \times n$ matrix of $(x, y)$ pairs in the regression data, if there were $n$ observations.

Using Bayes theorem, this can be turned into a probabilistic statement about $\theta$ given $D$ rather than the other way around:

$$P(\theta|D) = P(\theta)\frac{P(D|\theta)}{\int_\theta P(D|\theta)\, d\theta} \tag{2}$$

The expression $P(\theta)$ represents the Bayesian *prior* distribution, representing the analysts prior belief about the value of $\theta$ - often this is chosen to be *non-informative* - so that any value is equally likely before examining the data. For some models, the expression $\int_\theta P(D|\theta)\, d\theta$ is analytically soluable. However in general this is not the case. The advantage of Markov Chain Monte Carlo (MCMC) (see for example Gelfand and Smith (1990)) approaches to Bayesian analysis is that it works by simulating draws of random $\theta$ values from $P(\theta|D)$ - and that it does not require the integral expression mentioned earlier to do this.

In the situation described earlier, where there are some unobserved data items, let the unobserved data be denoted as $D^*$. If the conditional distribution of the unobserved data, given the observed data and the parameters is $P(D^*|D, \theta)$ then it can be noted that

$$P(D^*) \propto P(\theta)P(D|\theta)P(D^*|D, \theta) \tag{3}$$

Thus, provided an expression for $P(D^*|D, \theta)$ is known, posterior inference about the missing data $D^*$ can be made - and thus the stated aim of 'filling in the holes' may be ocheived. Again, this can be achieved via MCMC.

### 2.2 Applying Bayesian Methods to Time Series

As stated earlier, a full ARIMA model would require a large number of parameters - but in a situation with a relatively small number of observations this is impractical. For this reason, we use a simpler model - as set out below

$$\begin{aligned} P_t &= b_0 + b_1 t + b_2 t^2 + \varepsilon_t \\ \varepsilon_t &= \rho \varepsilon_{t-1} + \delta_t \\ \delta_t &\sim \mathrm{N}(0, \sigma^2) \end{aligned} \tag{4}$$

where $P_t$ is the population at time $t$ (here we denote 1990 as $t = 1$), and $\{\rho, \sigma, b_0, b_1, b_2\}$ are parameters to estimate - that is, together they constitute $\theta$. We assume $\delta_0 = 0$. This model can be thought of as a quadratic trend component (in some instances, $b_2$ is assumed to be zero, and a linear model is used), superimposed onto a random component. However the random component, rather than being independent for each value of $t$, is a random walk, so that the value of $\varepsilon_i$ is correlated to that for $\varepsilon_{i-1}$. Using MCMC methods, it is possible to estimate $\theta$, and also the missing observations, as suggested earlier.

## 3. Software Tools Used

Software for MCMC approaches has been, until recently, the province of the specialist. This altered with the release of *BUGS* (Lunn et al. 2009 2012) (Bayesian inference Using Gibbs Sampling). BUGS has now been extended with a Windows interface (WinBUGS) and to handle spatial data (GeoBUGS). However, data preparation, and post-modelling evaluation requires other software. The release of JAGS (Just Another Gibbs Sampler) (Plummer 2003) provides a further milestone. This offers a very similar facility to BUGS, but it is open source and may also be used in conjunction with the statistical programming language R via the *rjags* package in R. This offers R users the capability of fitting models using MCMC, but also exploits the power and flexibility of R, in order to prepare the data, and to provide extensive evaluation of the results. Using JAGS it is possible to obtain posterior distribution for the parameters outlined in the previous sections, and for the missing data. It is also worth noting that this approach also allows the constraint that the sum of all of the NUTS3 regions within a NUTS2 region must equal the statistic associate with the NUTS2 region.

## 4. A Basic Example

Below, a basic example of the approach is given, in this case based on the data for the Twente NUTS3 region in the Netherlands. Here, the actual data has no holes, and we intentionally deleted some observations. This way, the effectiveness of the approach can be assessed, as the 'true' missing value can be compared with the estimate. The result is illustrated in Figure 1. For each missing value, a point estimate is given, together with upper and lower credibility limits (a Bayesian equivalent of confidence intervals). In each case the true value is within the credibility limits, suggesting the method is effective. Note also that the technique identifies situations when there is a run of population values above the trend, below it, and also very close to it.

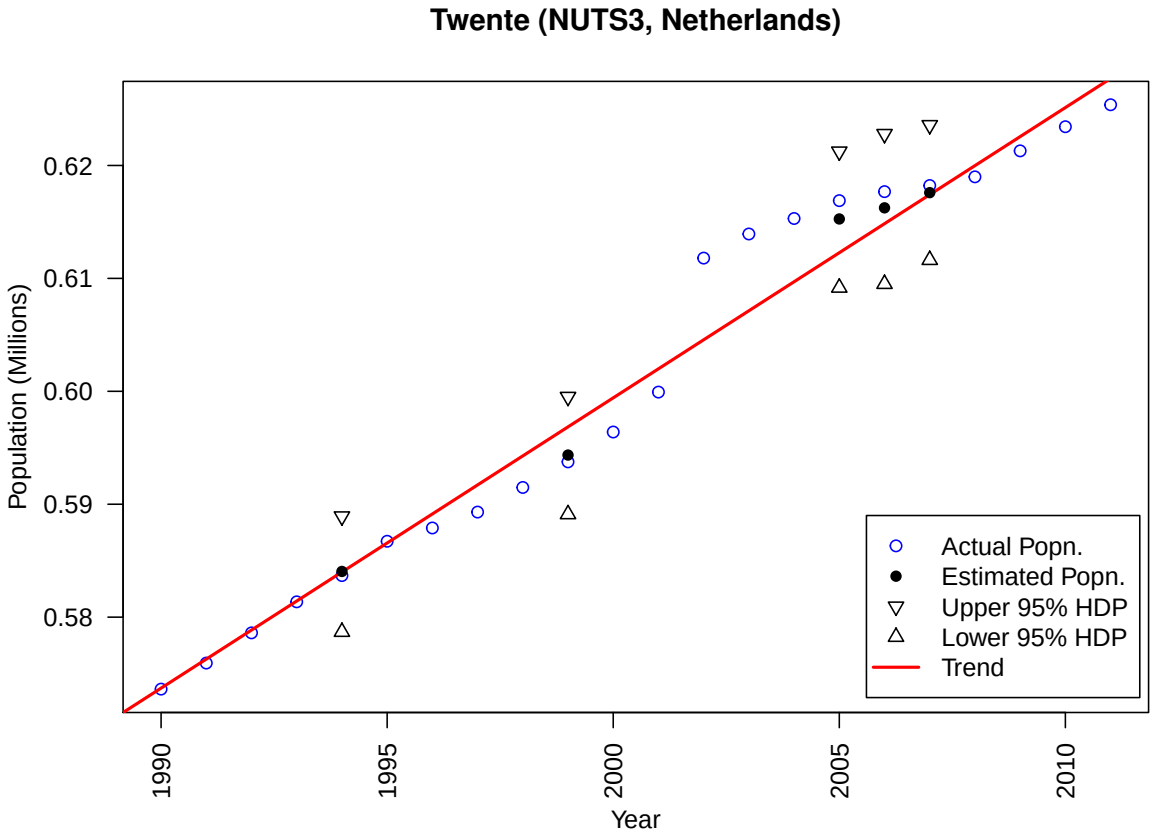



Figure 1: Estimates of missing populations, showing upper and lower 95% HPD limits.

## References

Chatfield, C. (1984). *The Analysis of Time series*. Chapman and Hall, London, 4th edition.

Enders, C. (2010). *Applied Missing Data Analysis*. Guilford Press, New York.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398—409.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS book: a practical introduction to Bayesian analysis*. Chapman and Hall, London.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The bugs project: evolution, critique and future directions. *Statistics in Medicine*, 29(25):3049–3067.

Plummer, M. (2003). JAGS: a program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna. ISSN 1609-395X.

# A Framework for Spatiotemporal Sensitivity Analysis of Geographical Models

Piotr Jankowski[1] , Arika Ligmann-Zielinska[2]

[1]San Diego State University, Department of Geography
San Diego, CA 92182-4493
pjankows@mail.sdsu.edu

[2]Michigan State University, Department of Geography
East Lancing, MI 48824-1117
ligmannz@msu.edu

## 1. Introduction

Uncertainty has been a prominent topic of inquiry in GIScience, transcending questions of representation, measurement accuracy, vagueness and ambiguity, all central to the subject of GIScience (Goodchild 2014, Fisher 2006). Much of the work on uncertainty has focused on positional error and its propagation in spatial data processing. A procedural framework presented here focuses on a systematic approach to analyzing attribute uncertainty arising from incomplete knowledge of model inputs and its influence on model results. The framework builds on the earlier work by the authors, presented at GIScience conferences (Ligmann-Zielinska and Jankowski 2008, Ligmann-Zielinska et al. 2012). Unlike previous contributions that focused on attribute uncertainty in spatial multiple criteria evaluation models, the framework presented here offers a comprehensive approach to analyzing attribute uncertainty in spatiotemporal models that are not limited to spatial multiple criteria evaluation.

## 2. Uncertainty vs. Sensitivity Analysis

Uncertainty analysis (UA) and sensitivity analysis (SA) are two methods of evaluating the uncertainty present in model input data. Since they both serve a similar purpose, they are often confused with each other and considered as interchangeable, whereas they should rather be used as two complementary methods. UA evaluates how the uncertainty of input data propagates through the model and affects its output values. This is a forward-chaining analysis of input uncertainty, resulting, conventionally, in statistical measures describing the variability of model output. UA does not explain what influence, if any, do model inputs have on the variability of model output. Explaining the influence and quantifying it for each specific model input is the function of SA. Consequently, SA is a backward-chaining approach that evaluates how much each source of input uncertainty contributes to model output variability. This complementary use of UA and SA, in which the former is a prerequisite to but not a substitute for the latter, is schematically depicted in Figure 1.

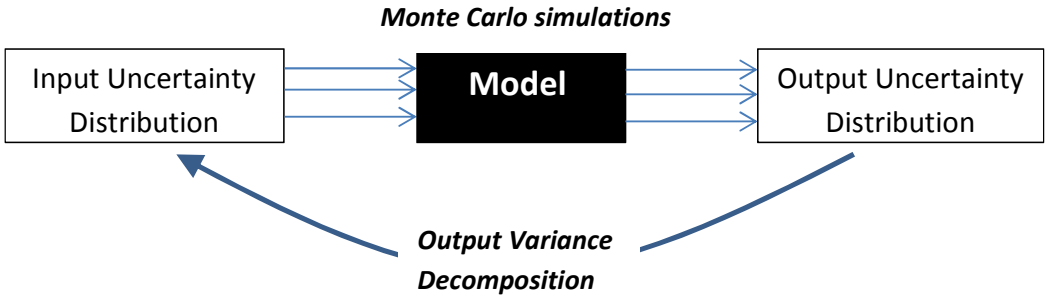



Figure 1. An integrated uncertainty-sensitivity analysis approach.

The idea underlying the integration of UA with SA (Figure 1) is based on the assumption that the uncertainty of a specific model input can be represented by a probability density function, which in turn can be sampled in the course of Monte Carlo process. In order to avoid sampling bias, an appropriately large sample of model's input values is used in a number of repeated model runs, resulting in a corresponding number of model outputs. A measure of central tendency (e.g. a mean) can be employed to capture a representative response of a given output element (e.g. a state variable), and its uncertainty (i.e. output uncertainty) can be represented by a measure of variability (e.g. a variance). This constitutes UA and is depicted by the left-to-right pass in the top part of Figure 1. The lower part of Figure 1, depicting the right-to-left pass, represents SA, in which variance of the model output is decomposed and apportioned it to the uncertain inputs, effectively expressing the (relative) share of each uncertain input in the model output variability.

## 3. The Method

We outline here specific steps of the integrated UA - SA method, which underpins the spatiotemporal sensitivity analysis framework. The steps are grouped into three phases: 1) simulating model outputs, 2) uncertainty analysis, and 3) sensitivity analysis.

**Phase 1: Model Output Simulation**

1. Identify the uncertain inputs in a model.
2. Generate a list of *N* input samples (parameter sets, or vectors of input variable values). The size of *N* depends on the complexity of a model.
3. Run the model *N* times to generate *N* realizations of the uncertain outcomes.
   3.1. If the model is temporal with *t* time steps, the total number of model executions amounts to $t \times N$.

**Phase 2: Uncertainty Analysis**

4. Aggregate the *N* model outputs:
   4.1. If the output is scalar, build its frequency distribution; for temporal output, produce *t* distributions.
   4.2. If the output takes form of spatial distribution (i.e. continuous or discrete), generate an *uncertainty map* (Ligmann-Zielinska et al., 2012).
   4.3. If the output is both spatial and temporal, generate for each time step *t* an *uncertainty map*.

**Phase 3: Sensitivity Analysis**

5. Use the UA results from phase 2 to perform output variance decomposition (Saltelli et al. 2010) and calculate:
   5.1. First-order sensitivity index ($S_i$) that quantifies the fractional contribution to outcome variance of a given factor *i* taken independently from other factors (Ligmann-Zielinska et al., 2012)..
   5.2. Time series of *Si* for time-dependent scalar (non-spatially distributed) output (Ligmann-Zielinska and Sun 2010).
   5.3. Generate maps of *Si* for spatially-dependent output (Ligmann-Zielinska and Jankowski 2014, Ligmann-Zielinska 2013).
   5.4. Animations of *Si* maps for spatiotemporal model output.
6. In order to examine input factor interactions and the nonlinear behavior of a model:
   6.1. Calculate total effect sensitivity index STi which quantifies the fractional contribution.
   6.2. Generate the map of factor dominance for Si and STi, respectively (Ligmann-Zielinska and Jankowski 2014).

## 4. The Framework

Given different representations of model outputs, the procedural framework covers four general cases, in which sensitivity analysis can be applied (Figure 2): [1] non-spatial and non-temporal (aggregate and static), [2] non-spatial and temporal (aggregate and dynamic), [3] spatial and non-temporal (spatially-dependent and static), and [4] spatiotemporal (spatially-dependent and dynamic). The graphical representations of the four types of sensitivity analysis progress from concise but low-information pie charts, through composite time series plots, through sensitivity maps depicting spatially-distributed influence of model inputs on spatially-distributed uncertainty of model output, to animations of sensitivity maps that are the most complex of all graphical representations proposed for the framework, but also carry the largest amount of cause-effect information about the particulars of model output uncertainty.

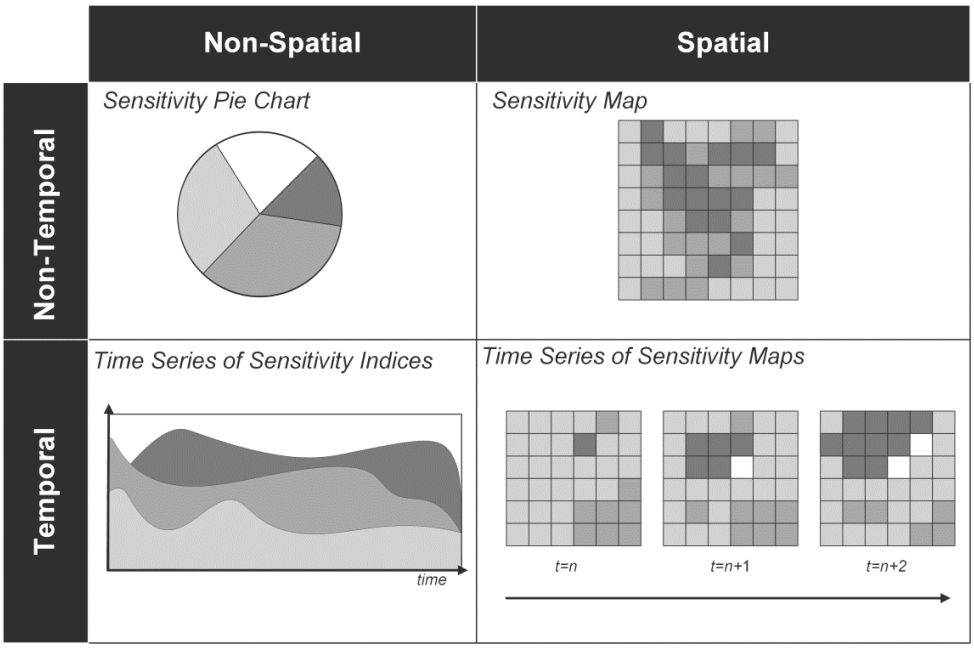


Figure 2. A framework for spatiotemporal sensitivity analysis and graphical representations sensitivity indicators.

The first case (non-spatial, non-temporal) fits models characterized by non-spatial (aggregated) inputs and outputs. An example can be a physical process model simulating total phosphorus (TP) concentration in lakes. The model, when used at a scale encompassing a system of lakes, uses aggregated (lumped) inputs and yields a single value of TP. Given that the model inputs are uncertain, the integrated uncertainty-sensitivity analysis, outlined in section 3, can be run resulting in calculating TP variance and sensitivity indexes.

The second case encompasses process models with spatially-invariant parameters. An example of such a model is an agent-based model (ABM), in which model inputs and outputs are aggregated (lumped) characteristics of space, which change with each time step (i.e. model simulation run) (Ligmann-Zielinska and Sun 2010). The integrated uncertainty-sensitivity analysis can be repeated for each simulation run resulting in trajectories, depicted on time-series plots, showing the change in sensitivities of model inputs over time.

The third case corresponds to models, in which model inputs and outputs are spatially distributed (e.g. represented by raster or vector layers) and the model output is static (i.e. time-invariant). An example of such a model is a multiple-criteria land suitability evaluation model (Ligmann-Zielinska and Jankowski 2014). In this case, the variance of land suitability index (model's output) and sensitivity values of input criteria can be calculated and mapped for each spatial entity (i.e. a raster cell or a polygon) within the modeled area.

The fourth case, the most complex of all, refers to spatially-explicit process-based models, in which the spatially-explicit uncertainty-sensitivity analysis is repeated at each time step of model simulation.

All four cases apply to uncertainty and sensitivity analysis of parameter values (e.g. weights assigned to layers) and to feature attribute values (i.e. spatial variable values). It should be acknowledged that the variance decomposition-based uncertainty and sensitivity analysis of feature attribute values can be computationally challenging given a large number of analyzed features.

## 5. Summary and Conclusions

The presented framework for SA accounts for spatial and temporal dimensions of geographic models. We identify four classes of SA, which serve different purposes. In particular, the framework can be utilized to identify model inputs responsible for static model outcome variability measured over the entire area (non-spatial and non-temporal), to evaluate the temporal sensitivity of system-wide outcomes due to changes in model inputs (non-spatial and temporal), to identify spatial clusters of high model sensitivity to particular inputs - regardless of time (spatial and non-temporal), and to comprehensively evaluate how spatial clusters of high model sensitivity evolve over time (spatiotemporal).

Understanding and categorizing model output uncertainty through sensitivity analysis is critical if one wants to use geographic models to address complex spatially heterogeneous problems including urban expansion, climate change, deforestation, disease spread, and water pollution. An inherent complexity in these processes, leading to model uncertainty, is the

variability over space and time. The presented framework introduces a comprehensive approach to SA allowing for uncertainty evaluation in various types of spatial models.

## ACKNOWLEDGMENTS

Partial support for this research was provided by the National Science Foundation Geography and Spatial Sciences Program Grant No. BCS 1263477. Any opinion, findings, conclusions, and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

Fisher, P. F., Ed. (2006) *Classics from IJGIS: Twenty Years of the International Journal of Geographical Information Science*. Boca Raton: CRC Press.

Goodchild, M. F. (2014). Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science*, (1), 3–20.

Ligmann-Zielinska, A. 2013. Spatially-Explicit Sensitivity Analysis of an Agent-Based Model of Land Use Change. *International Journal of Geographical Information Science*, 27(9): 1764-1781. \

Ligmann-Zielinska, A. and P Jankowski. 2014. Spatially-Explicit Integrated Uncertainty and Sensitivity Analysis of Criteria Weights in Multicriteria Land Suitability Evaluation. *Environmental Modelling & Software*, 57, 235-247.

Ligmann-Zielinska, A., P Jankowski, and J Watkins. 2012. Spatial Uncertainty and Sensitivity Analysis for Multiple Criteria Land Suitability Evaluation. *Extended Abstract, Seventh International Conference on Geographic Information Science*. Columbus, OH, U.S., September 18-21, 2012.

Ligmann-Zielinska, A. and L. Sun. 2010. Applying Time Dependent Variance-Based Global Sensitivity Analysis to Represent the Dynamics of an Agent-Based Model of Land Use Change. *International Journal of Geographical Information Science* 24:1829-1850.

Ligmann-Zielinska A and Jankowski P, 2008, A framework for sensitivity analysis in spatial multiple criteria evaluation. In: Cova TJ, Miller HJ, Beard K, Frank AU and Goodchild MF (eds), *Geographic Information Science Proceedings 5th International Conference*, *GIScience 2008*, 217-233. Berlin/Heidelberg: Springer.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S. 2010. Variance based sensitivity analysis of model output: Design and estimator for the total sensitivity index. *Computer Physics Communications*. 181: 259-270.

# Shade Optimization in a Desert Environment

Qunshan Zhao[1], Elizabeth A. Wentz[1], Alan T. Murray[1]

[1]GeoDa Center for Geospatial Analysis and Computation, School of Geographic Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287-5302, USA
Email: {qszhao; wentz, atmurray}@asu.edu

## 1. Introduction

Shade provided by trees, shrubs and other natural vegetation serves as a natural umbrella for residential and commercial buildings. In desert regions like Tempe, Arizona strategically located shade can translate into significant energy and long-term cost savings as well as prove beneficial to human health and well-being. Our research goal is to develop a process for placing trees strategically to maximize shade coverage on 3D surface for residential structures.

We address two optimization challenges. One is that tree shade geometry varies by species, height, age, and location, making optimal placement a spatially dependent problem. While classic modelling approaches exist to support spatial coverage optimization, the irregular tree shade coverage provided to building structures means that analysis and manipulation are not straightforward. Research addressing irregular shapes suggests assuming a particular geometry, such as "S" shapes (Hof & Joyce 1992), but this is not sufficient for general application. The second challenge is that maximizing coverage of three-dimensional (3D) structures requires differentiation relative to walls, windows and doors, and rooftops. Research on 3D optimizing coverage has relied processing and interpretation that is essentially two-dimensional (2D). Goodchild & Lee (1989) formulate a visibility coverage model based on the location set covering problem (LSCP) and the maximal covering location problem (MCLP) to minimize the number of viewpoints and maximize the visible areas on an irregular topographic surface. Murray et al. (2007) model security sensor placement in 3D urban environments by using the MCLP and the backup coverage location problem (BCLP). In both cases, 3D visibility assessment is translated into 2D coverage, enabling these classic location coverage models to be applied once visibility analysis is carried out. While some similarities in approach and evaluation can be made to tree shade optimization, existing approaches do not contend with the intricacies of 3D coverage provided to structures across space.

In general, the goal of our research is to develop an approach for spatial optimization on 3D surface by GIScience methods. As for this example, we attempt to decide the best tree placement that maximizes shade coverage of a single-family house. This will enable development of landscaping and building structure design guidelines that mitigate direct sunlight and heat intensity. The study area is a residential neighborhood in Tempe, Arizona. Quickbird imagery and LIDAR elevation data are used to identify and model the 3D geometry of trees and single-family homes.

## 2. Spatial optimization

The basic premise of this research is that quantitative methods can be brought to bear on an important issue in a desert environment, that of achieving greater energy efficiency and mitigating heat related illness and deaths. To accomplish this, a spatial optimization model is developed and applied. Given space limitations here, a generic spatial optimization model is used to describe the basic approach based on Tong & Murray (2012):

$$Maximize \quad g(x) \tag{1}$$

$$Subject\ to \quad f_i\ (x) = b_i \ \ \forall \mathrm{i} \tag{2}$$

$$x\ binary \tag{3}$$

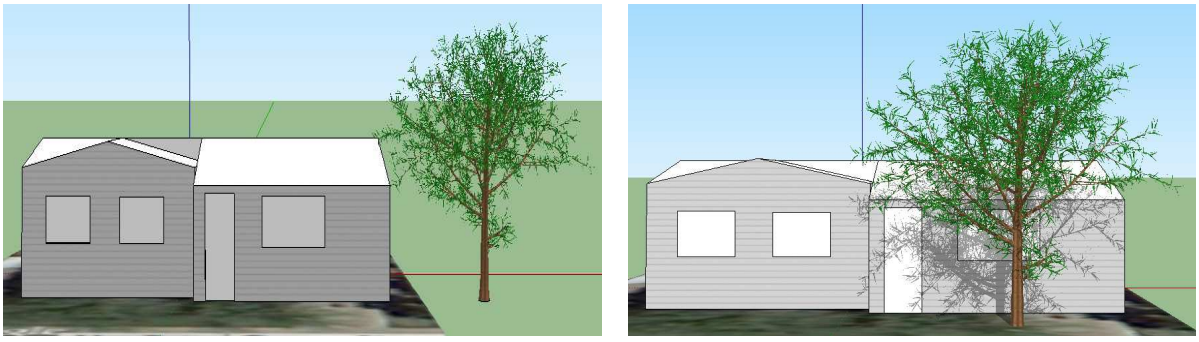
where x is a vector of decision variables associated with where to locate one or more trees, $x = [x_1, x_2, x_3, \ldots x_n]^T$, $g(\,)$ and $f_i(\,)$ are functions, and $b_i$ is a coefficient for each constraining condition $i$.

The details of the particular problem examined here vary from this generic formulation, but the fundamental features can be summarized as follows. The objective, (1), is a function specifying the shading of structures in a 3D urban environment, and this is to be maximized. The constraints, (2), relate tree siting to shade coverage provided. Finally, constraints (3) indicate that the decision variables *x,* binary requirements.

## 3. Research methods

Our study area is in Tempe, Arizona, a municipality in the greater Phoenix Arizona metropolitan area with approximately 160,000 residents. Summer temperatures in Tempe can exceed 43 °C, requiring residents to rely on heat mitigation strategies such as home air conditioning, swimming pools, and shade.

The data for our project includes a number of different inputs. Quickbird imagery is used to classify objects as trees, shrubs, grass, impervious surface, water (swimming pools), and buildings. LIDAR elevation data is used to refine our classification system and provide building and tree height information. Figure 1a illustrates a typical residential structure and a tree.[1] The house is a one-story ranch home that is roughly 150 $m^2$. The tree is a 9-meter high thornless mesquite (*Prosopis thornless hybrid 'AZT™'*). The goal of the work is to find a better location for the tree in terms of shade provision, such as that shown in Figure 1b where the shade is cast for 12 PM (noon) on December 22nd.



(a) Initial tree location (b) Enhanced shade through tree relocation

Figure 1. Modelling results for a single-family residence and a thornless mesquite tree.

[1] Building and tree structures created using Google map plugin and 3D tree plugin in Google Sketchup

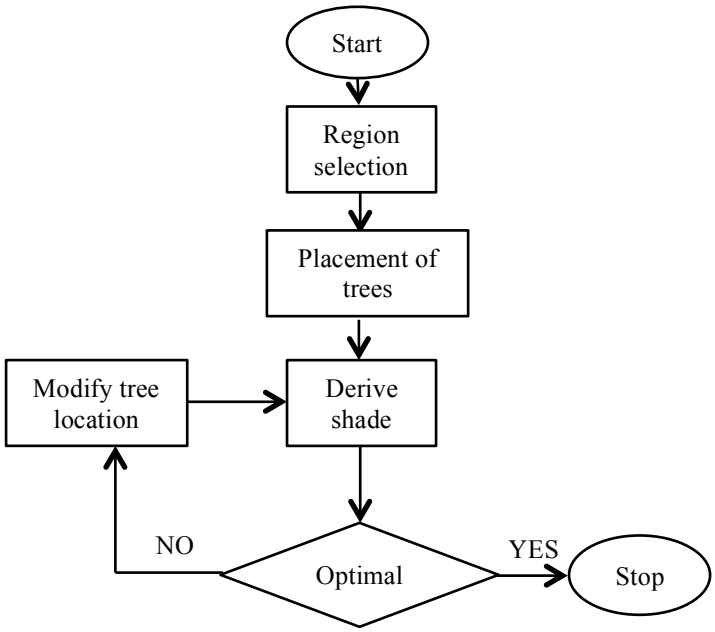


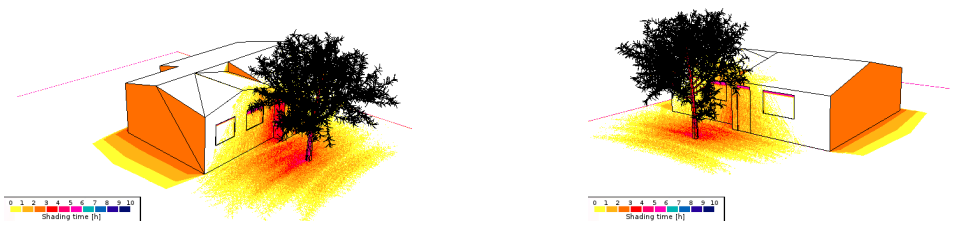

Figure 2. Landscape design approach to enhance shade provision.

The primary components of the approach to optimize shade are shown in Figure 2. Technically, the model formulation given previously reflects the problem of where to locate a single tree. The criteria associated with measuring shade effects on different parts of a single-family residence is based on the work of Shaviv & Yezioro (1997), who proposed using a geometrical shading coefficient (GSC) to express the ratio between shaded and total examined surface areas.

## 4. Preliminary results

While conventional wisdom suggests tree placement on the south and west of buildings (in the northern hemisphere), we calculate the shade value as the shadow footprint moves from west to east during the day. This involves placing a single tree and deriving shade during six discrete time periods (July 15 at 10am, 11am, 12pm, 1pm, 2pm and 3pm). These periods of the day represent the greatest heating potential.

The spatial optimization problem ranks the importance of potential tree location areas by shadow benefits and landscape design constraints. For shadow benefits, windows and doors have priority for coverage, followed by house walls, and the roof. We also restrict tree placement at 3.5-4.5 meters from the house wall due to the growth needs of the 9-meter thornless mesquite. Using this approach, the optimal tree location for a residential structure is identified, and is shown in Figure 3.



(a) West facing view (b) East facing view

Figure 3. Optimal shade coverage intensity.

Figure 4 reports the GSC value for the door and windows, walls and roofs during the 6 hourly time periods between 10 am and 3 pm. Results show that windows and door obtain more than 20% of the shade coverage during this time period. The south walls are covered 10% by shade during this time. Roof coverage is comparatively small because of the high solar elevation angle during the summer months and the distance between the tree and the house.

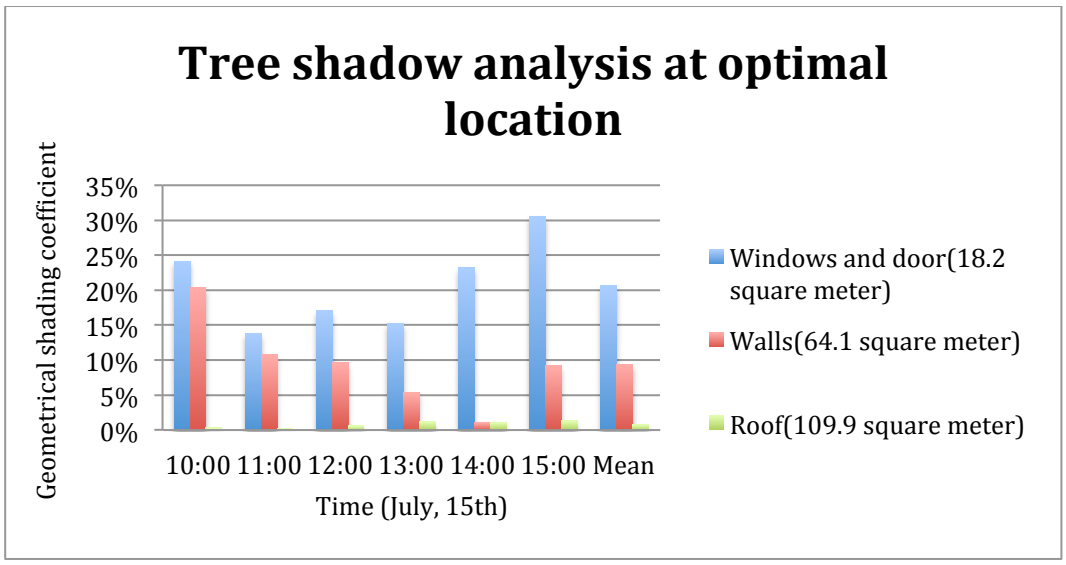



Figure 4. Shadow analyses at optimal location.

## 5. Future work and conclusion

The preliminary results show shadow effects for a single-family house. We show how to differentiate spatial coverage on a 3D structure by calculating the shadow coverage on house facades and the roof. Future work will extend this research to include a larger geographic study area, multiple houses, multiple trees and varying tree species.

## Acknowledgements

The authors want to thank the Decision Center for a Desert City at Arizona State University for providing remote sensing data and support for this research.

## References

Goodchild, M. F., & Lee, J. (1989). Coverage problems and visibility regions on topographic surfaces. Annals of Operations Research, (18), 175–186.

Hof, John G., & Joyce, Linda A. (1992). Spatial Optimization for Wildlife and Timber in Managed Forest Ecosystems. Forest Science, (38), 489-508.

Murray, A. T., Kim, K., Davis, J. W., Machiraju, R., & Parent, R. (2007). Coverage optimization to support security monitoring. Computers, Environment and Urban Systems, 31(2), 133–147. doi:10.1016/j.compenvurbsys.2006.06.002

Shaviv, E., & Yezioro, A. (1997). Analyzing mutual shading among buildings. Solar Energy, 59, 83–88.

Tong, D., & Murray, A. T. (2012). Spatial Optimization in Geography. Annals of the Association of American Geographers, 102(6), 1290–1309. doi:10.1080/00045608.2012.685044

# Pattern-based approach to knowledge extraction from giga-cell geospatial raster datasets

J. Jasiewicz[1], P. Netzel[2], T. F. Stepinski[3]

[1]Adam Mickiewicz University, Dziegielowa 27, 60-680 Poznań
Email: jarekj@amu.edu.pl

[2]University of Wroclaw, Kosiby 6/8, 51-621 Wrocław, Poland
Email: pawel.netzel@uni.wroc.pl

[3]University of Cincinnati, Cincinnati, OH 45221, USA
Email: stepintz@uc.edu

## 1. Introduction

There is a keen interest in development of automated methods of knowledge discovery from large geospatial raster datasets. A geospatial raster is large either because it has an extensive geographical coverage or because it has a very fine resolution. Standard GIS methods of analysis are ill-suited for such datasets as they operate either at the level of individual raster cell or at the level of multi-cell "object" (Blaschke, 2010). Such approaches will provide little insight in application to large (say, having $10^9$ cells) raster. This is because the size of the cell and the data spatial extent differ by too many orders of magnitude. To interpret and analyze giga-cell rasters we propose an addition to the present GIS paradigms – a spatial analysis of local patterns. A local pattern (a "scene") is a mosaic of raster cell values. Importance of spatial patterns is in their association with specific geographic notions. For example, in a land cover dataset, a specific pattern of land cover categories can be recognized as a "downtown area." At smaller scale, an analyst could be interested in an actual configuration and composition of land cover categories constituting downtown, but at the large scale it's sufficient and indeed necessary to interpret this location by a single label. Traditionally, pattern analysis has been a domain of computer vision (Datta et al., 2008), whereas spatial analysis has been a domain of GIS. We have integrated the two domains by developing a methodology that enables GIS processing of local patterns.

## 2. Pattern-based GIS analysis of rasters

### 2.1 Defining scenes

Scenes are extracted from a raster in three different manners (see Fig.1B). (a) From neighbourhoods around a given set of points. (b) From a set of irregular segments. (c) As a regular grid-of-scenes. A grid-of-scenes covers the extent of original raster but has a much larger cell size. Grid-of-scenes stores pattern information extracted from square scenes centered on its cells; scenes overlap if they are larger than grid-of-scenes cells. A scene is represented by a pattern signature which encapsulates its character. Signatures are stored in the grid-of-scenes. We use histograms of pattern features as signatures. Pattern features are pieces of information about a pattern; only categorical features can be histogramed so numerical features must be categorized before they can be used. No single set of features describes well all possible patterns. GeoPAT – the software implementation of our methodology – implements several different methods of feature extractions.

### 2.2 Comparing scenes

In order to conduct operations on scenes we define a "distance" between the two scenes to be a distance between histograms of their features. There is no a single, universally accepted measure of distance between two histograms. Moreover, distance function needs to be matched with features selected for scene signature. In GeoPAT we have implemented several distance functions that match well with types of features we use. For example, to quantify patterns of land cover we use two

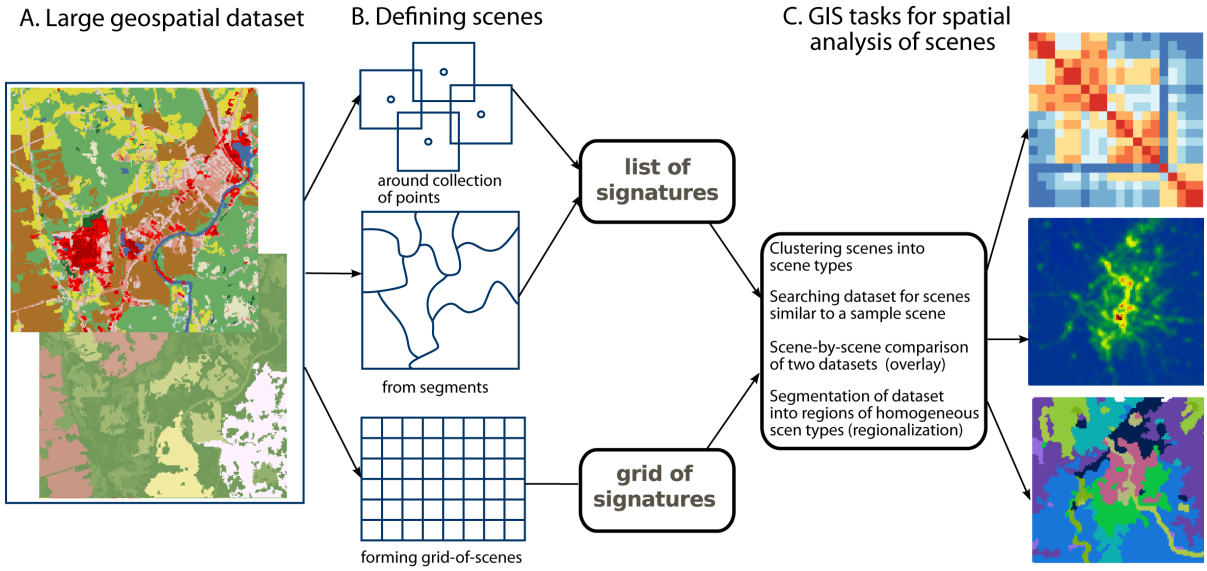



Figure 1: Application of pattern-based analysis to 30m/cell DEM covering the entire extent of the country of Poland.

features (land cover category and categorized size of clump to which a given cell belongs (Jasiewicz and Stepinski, 2013a) and we measure distance between histograms using Jensen-Shannon divergence (Lin, 1991). To quantify topographic patterns we use co-occurrence features (Haralick et al., 1973) and the Wave Hedges distance.

### 2.3 GIS processing of patterns

GIS tasks are performed on the grid-of-scenes with a scene signature serving as a new type of cell attribute and a query-by-pattern-similarity (QBPS) based on histogram distance replacing the SQL. Specifically, we have implemented four GIS tasks for spatial analysis of scenes (see Fig.1C). (a) Clustering of scene collection (for example, into landscape types or physiographic types). (b) Search of dataset for scenes similar to a given sample pattern with results visualized in terms of similarity map that reveals a geospatial context of the sample pattern. (c) Comparison between two co-registered grids-of-scenes which is equivalent to a standard GIS overlay function and can be used for assessing change in local patterns. (d) Segmentation of grid-of-scenes into sub-regions of uniform patterns. This task performs regionalization of a raster with respect to patterns of original variable. For example, if applied to a land cover dataset it will map its constituent landscape types, and, when applied to a topographic data, it will map its physiographic regions.

## 3. Application to topographic dataset of Poland

To demonstrate our methodology we have performed all four tasks described in section 2.2 using a 30m/cell DEM (21,696 × 24,692 cells) covering the territory of Poland. First, the DEM was categorized into 10 landform elements using the geomorphons method (Jasiewicz and Stepinski, 2013b). The landform elements map is shown in the center of Fig.2. We constructed a grid-of-cells with the resolution of 1.5 km (50 times larger than a resolution of the DEM) and extracted (overlapping) scenes having size of 15×15 km each. We used co-occurrence features to construct histograms and the Wave Hedges function to measure distances between histograms. Panel A shows schematically the results of the clustering task. A collection of scenes is clustered using hierarchical clustering method (shown here graphically in a form of a “heat map”). In this example the collection splits into three clusters interpreted as mountains, hills, and lowlands, respectively. Panel B shows the working of the search task. Three queries are shown each resulting in a creation of similarity map visualizing spatial extent of terrain similar to the corresponding query. Panel C shows a comparison between two rasters. In this case both rasters are categorizations of the original DEM but using different parameters, the result is a map showing regions where change in categorization parameters

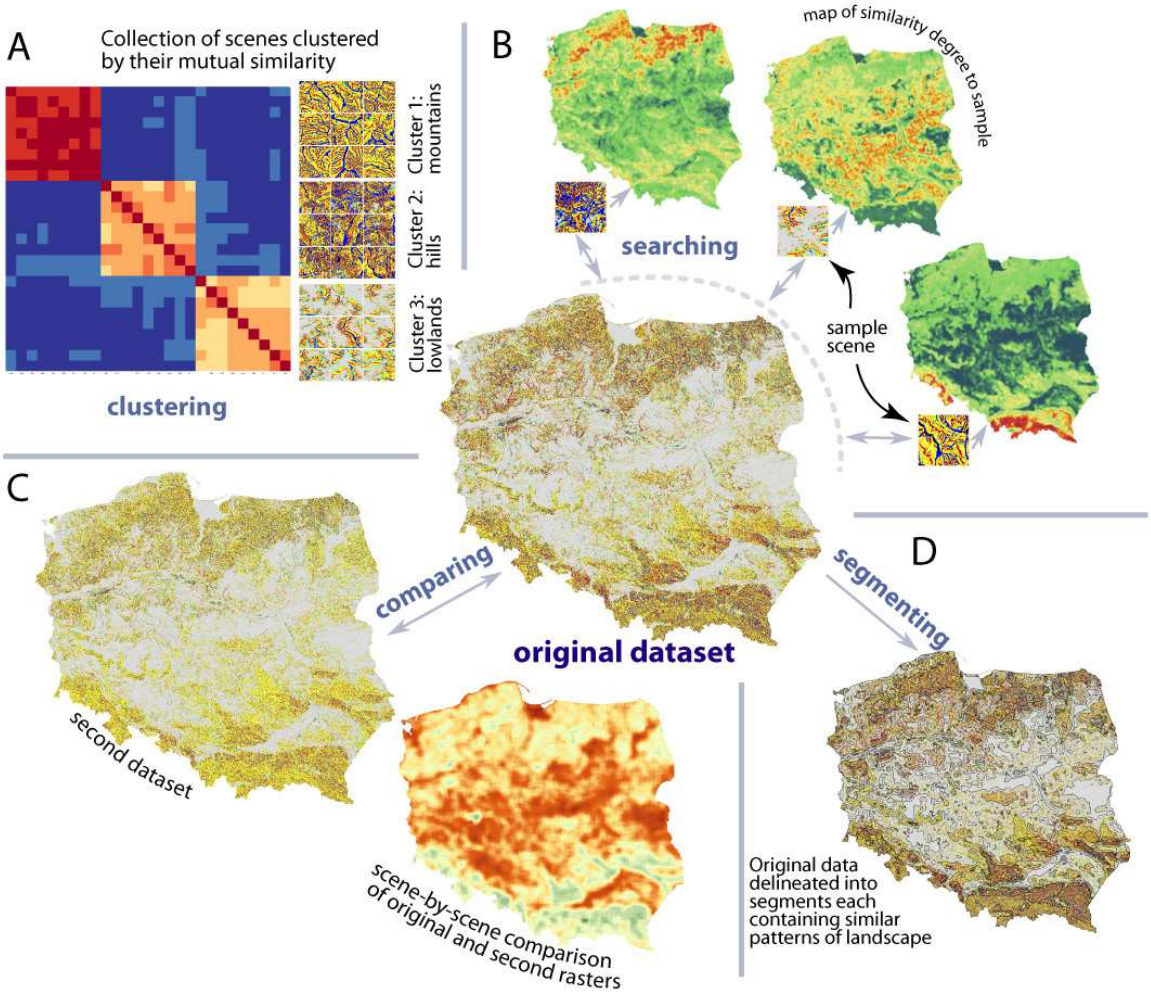



Figure 2: Application of pattern-based analysis to 30m/cell DEM covering the entire extent of the country of Poland.

results in significantly different interpretations of landscapes. Panel D shows results of the segmentation task. The territory of Poland is divided into irregular segments so that topographic pattern (landscape) in each segment is uniform. Classification of those segments yields a physiographic map of Poland (see Jasiewicz et al., 2014 for details). Such map could be thought of as the result of unsupervised machine learning. Alternatively, by selecting samples of major physiographic regions in Poland and running search task, a supervised learning-based physiographic map of Poland can be created as well.

## 3. Conclusions

We presented a methodology for knowledge extraction from giga-cell raster datasets. The methodology is based on spatial processing of local patterns formed by an original dataset variable. Computer visions concepts of pattern representation and similarity are used to extend several GIS tasks to work with grids of patterns. All steps necessary to apply our methodology in practice are contained in the toolbox GeoPAT which is freely available for download from http://sil.uc.edu/gitlist . This toolbox is a collection of newly written GRASS GIS modules. Integration with GRASS assures that very large datasets can be processed with ease and makes possible creation of scripts for computational pipelines that use other GRASS, as well as R modules. Such integration make complex tasks, such as, for example, supervised or unsupervised delineation of physiographic regions over large extents, relatively fast and easy. In addition to land cover and topography, our method can be applied to multiple other domains including crops, soils, climate, phenology, as well as socio-

economic data. It can also be applied to very large, high resolution images where multiple instance learning approach (Vatsavai, 2013) is currently used for the segmentation task.

## Acknowledgements

The work was supported by NSC GRANT DEC-2012/07/B/ST6/01206, by the NSF under Grant BCS- 621 1147702, and by the University of Cincinnati Space Exploration Institute.

## References

Blaschke, T., 2010, Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens*. 65(1), 2–16.

Datta, R., Joshi, D., Li, J., Wang, J. Z., Apr. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Comput. Surv. 40 (2), 1–60.

Jasiewicz, J., Stepinski, T. F., 2013a. Example-Based Retrieval of Alike Land-Cover Scenes From NLCS2006 Database. *IEEE Geosci. Remote Sens. Lett.* 10 (1), 155–159.

Jasiewicz, J., Stepinski, T.F., 2013b. Geomorphons-a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156.

Jasiewicz, J., Netzel, P., Stepinski, T.F. 2014. Landscape similarity, retrieval, and machine mapping of physiographic units. Geomorphology 221, 104–112.

Haralick, R. M., Shanmugam, K., Dinstein, I., Nov. 1973. Textural features for image classification. Syst. Man Cybern. IEEE Trans. 3 (6), 610–621.

Lin, J., 1991. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 31(1), 145–151.

Vatsavai, R.R. 2013. Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery. In Proceedings of the 19th ACM SIGKDD, 1419-1426.

SAM: A Provenance Model for Spatial Analytical Methods

Wenwen Li, Sergio Rey, Luc Anselin

GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning
Arizona State University
Tempe AZ 85287-5302
USA
Email: {wenwen, srey, Luc.Anselin}@asu.edu

# 1. Introduction

Provenance, which captures and records the history and source of a dataset or a process, is a critical topic in many domains for evaluating the quality and validity of scientific output (Miles et al. 2007). In a field where many data- and compute- intensive applications are seen, such as GIScience, it is of great importance to understand how a dataset is produced, which method is applied, what computing platform the process runs on, and who is the person or organization responsible for the dataset. These elements are all part of provenance information and can greatly facilitate: (1) the evaluation of data reusability; (2) detection of potential errors in the final result by backtracking previous processing steps; (3) easy maintenance of copyright and ownership of data; and more importantly, (4) replicability of scientific research.

It is useful to draw a distinction between data provenance and provenance for spatial analytical methods (Anselin et al. 2014). The former centers on data resources and recording data lineage in form of attribution metadata, similar to the FGDC (Federal Geography Data Committee) metadata associated with an Esri shapefile. The latter focuses on a spatial analytical process and defines the source data it uses, the method in use and the results that are generated. Though both provenance types describe data and related processes, data provenance is most often used to help a user understand a dataset and evaluate its usage, whereas the provenance for spatial analytical methods is tailored to communicate with other processes which are part of a complex processing chain or scientific workflow. The data in this case, is only the medium that flows along these processes and gets modified.

Data provenance research has received much attention in and beyond the GIScience community (Buneman et al. 2000; Tilmes et al. 2010; Bennett et al. 2011). However, provenance for spatial analytical processes remains largely in its infancy. Most research remains at the conceptual level with two notable exceptions. Di et al. (2013) proposes the use of International Standardization Organization (ISO) 19115:2003 (ISO 2003) and ISO 19115-2:2009 (extension for gridded and imagery data; ISO 2009) to record the provenance information captured during a geoprocessing workflow. This application and geoprocessing methods, as well as the ISO 19115 lineage models, however, are more suitable for raster data processing. Anselin et al. (2014) proposes a new metadata structure – WMD (Weights MetaData) to capture the

provenance in spatial weights generation, one of the fundamental components in many spatial analysis methods. This new structure uses lightweight JSON (JavaScript Object Notation) to encode the provenance information during the execution of spatial weights operations. Besides provenance capture, a provenance-tracing algorithm is developed to recursively trace the processing steps for a weights dataset and has the ability to reproduce the results according to what is captured in the provenance model automatically.

In this paper, we built upon our earlier successful experience on spatial weights metadata to propose a new provenance model SAM (Spatial Analytical Metadata). SAM is envisaged as both a general and flexible framework that can be used by a variety of spatial analysis methods and workflows. We provide an illustration of the framework involving two examples, one in spatial weights and the other in spatial autocorrelation analysis to demonstrate the applicability of this new metadata to capture the provenance in spatial analytical methods. We also develop a mapping between the SAM model and the widely adopted W3C PROV model to demonstrate its interoperability.

# 2. The SAM model for spatial analysis

Figure 1 demonstrates the hierarchical structure of the SAM model for encoding provenance information of a spatial analytical operation. The nodes highlighted in black borders are shared with the WMD model. They include the input data and output data in the desired format and the location represented using a URI (Universal Resource Identifier). URI refers to a unique address of a dataset. It could be located on a file system on a local machine (with prefix “file://”), or within a Web accessible folder (with prefix “http://”). One unique feature of this model is that it not only allows input data to be directly given, but also allows the input data type as a derived dataset. In this case, the URI of the dataset is not given, rather the process procedure of the data is specified in a WMD or SAM file. Another shared node between SAM and WMD is labeled as “parameters”, recording the runtime parameters for different spatial analysis methods. For example, to generate spatial weights based on a kernel density, the runtime parameters include the number of nearest neighbors (*k*) to use in determining the bandwidth and the specific kernel function (function), such as Gaussian or uniform.

The SAM model also contains extensions of the WMD model (pink nodes). These extensions provide auxiliary information about the execution of a spatial analytical method. For instance, “software” and “algorithm” defines the tool and algorithm used to perform some scientific workflow. The illustration refers to specific elements of PySAL (Rey and Anselin, 2007). “Platform” records the computing environment on which the procedure runs; it could be a single desktop or in a cloud computing environment. This information will be helpful to evaluate/standardize the runtime used for execution, given in the nodes of “CreationTime” and “FinishTime”. Additionally, a node “person” is created in SAM to record the responsible party who creates or modifies a dataset by performing a specific type of spatial analysis operations.

The node “analysis_type”, is a generalization and is different from that (“weight_type”) defined in the WMD model to allow provenance recording of multiple types of spatial analytical methods. For example, if a spatial autocorrelation method, such as Global Moran’s I’s provenance is to be captured, the value for the “analysis_type” key would be “Moran”, the input data would be a spatial weight file and the attribute table; the runtime parameter will include the attribute name for which Moran’s I will be performed. Figure 2 demonstrates such an example. The two input datasets are a derived weights data and an attribute data in text format. The output is a floating number “p_normal” to help determine if the data has shown significant pattern of spatial autocorrelation. All the auxiliary information, though will not impact the execution process, provide important runtime information for the process.

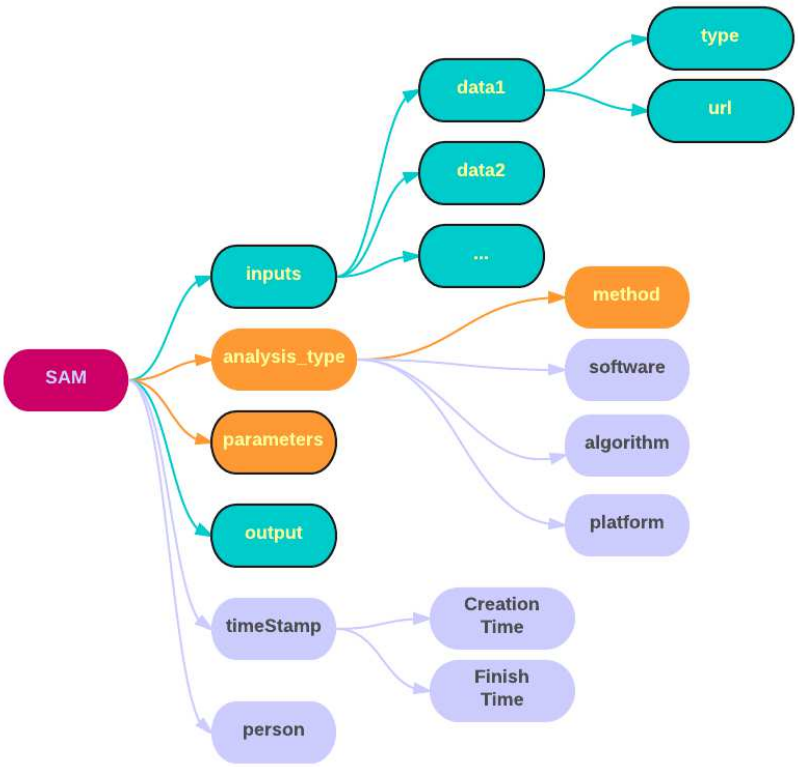



Figure 1. Metadata structure of SAM model. Green nodes refer to input and output data. Orange nodes are spatial analysis operations and runtime parameters. Purple nodes include auxiliary information for the process.

```
{
  "input":{
    "weights": {
          "type": "prov",
           "uri": "http://toae.org/pub/taz_block.wmd"},
    "attribute": {
          "filetype": "txt",
          "name": "HR7984",
          "uri": "http://toae.org/pub/stl_home.txt"}
        },
    "analysis_type": {
          "method": "Moran",
          "software": "PySAL",
          "algorithm": http://pysal.readthedocs.org/en/v1.7/library/esda/moran.html",
          "platform": "single desktop"}
    "parameters": {
          "transform": "O"},
    "output": {
          "p_norm": "float"},
    "timestamp": {
          "creationTime": "",
          "software": "PySAL"}
    "person": "ASUGeoDaTeam"
}
```

Figure 2. An example of provenance structure for spatial autocorrelation

# 3. Mapping the SAM model to the W3C PROV model

We introduced the SAM model and ISO 19115, with the former focusing on provenance of spatial analytical method and the latter focusing on describing raster data provenance. In fact, there is another wide-adopted provenance model - PROV, produced by World Wide Web Consortium (W3C) to support the definition and interoperable exchange of provenance information across heterogeneous platforms. PROV describes the provenance information by a triple <agent, activity, entity>. As an upper-level provenance model, PROV is general enough to represent either data or analysis provenance. In this paper, we map the proposed SAM model to PROV to demonstrate its characteristics of being highly interoperable.

Figure 3 demonstrates a PROV model that maps all provenance elements in the SAM model. The different colors and shapes represent elements of agent, entity and activity respectively. Each role (arrow) matches the provenance key in SAM and the end node of each role is the actual value for that particular key. For instance, when an operation (defined in “Analysis_Type”) is initiated, a corresponding software tool (node “Software”) must be invoked. Therefore, the relationship between the two nodes are “wasStartedBy” with namespace “prov”.

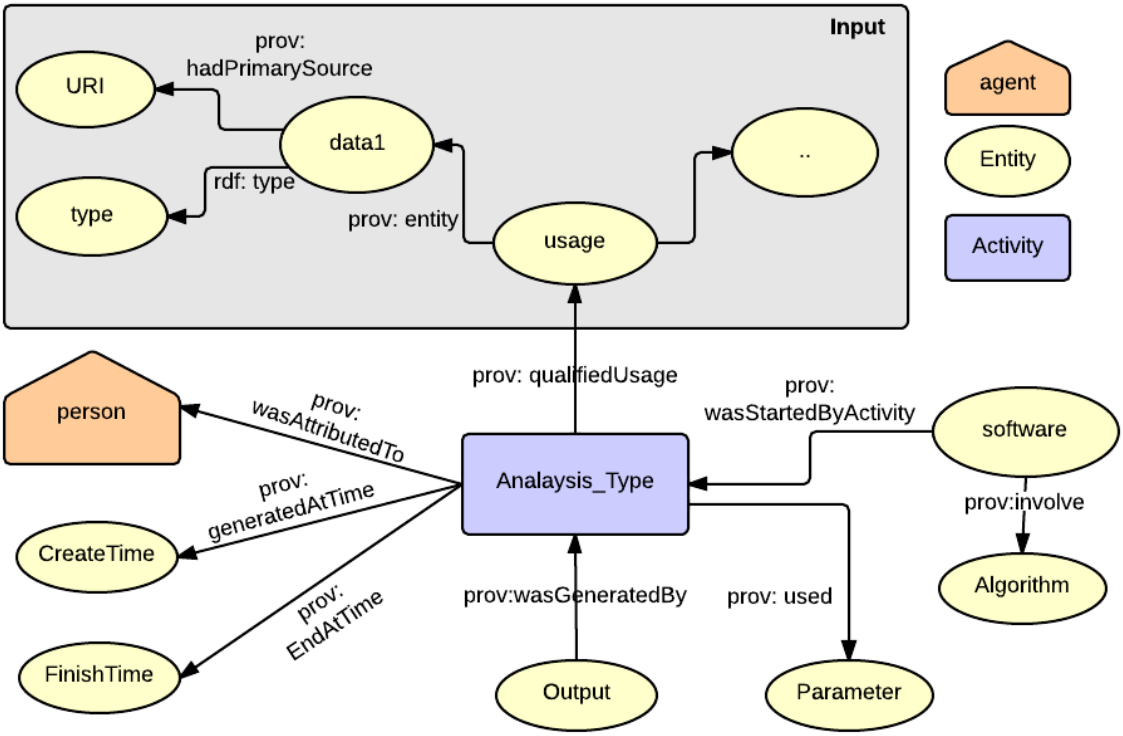



Figure 3. Mapping SAM to PROV

# 4. Conclusion and discussions

This paper introduces a new provenance model - SAM for capturing provenance of spatial analysis operations. Besides inheriting from the WMD model in its light-weighted data structure, its ability to enable automated provenance tracking, SAM can also be nicely mapped to other existing provenance models, such as W3C PROV, demonstrating its advantage in enabling provenance interoperability.

# Acknowledgements

Funding from the National Science Foundation (NSF SI2-SSI: CyberGIS Software Integration for Sustained Geospatial Innovation) is gratefully acknowledged.

# References

Anselin L, Rey S and Li W, 2014. Metadata and provenance for spatial analysis: the case of spatial weights. *International Journal of Geographic Information Science*. (forthcoming)

Bennett DA, Tang W and Wang S, 2011. Toward an understanding of provenance in complex land use dynamics. Journal of Land Use Science, 6(2-3): 211-230.

Buneman P, Khanna S and Tan WC, 2000. Data provenance: Some basic issues. In FST TCS 2000: *Foundations of Software Technology and Theoretical Computer Science* (pp. 87-93). Springer, Berlin Heidelberg.

Miles S, Wong SC, Fang W, Groth P, Zauner KP and Moreau L, 2007. Provenance-based validation of e-science experiments. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1): 28-38.

Rey, S.J. and L. Anselin, (2007) PySAL: A Python library of spatial analytical methods. *The Review of Regional Studies*, 37(1):5–27.

Tilmes C, Yesha Y and Halem M., 2010. Tracking provenance of earth science data. *Earth Science Informatics*, 3(1-2): 59-65.

# Characterizing relationships between data aggregation and spatial scale: Exploratory analysis of the Modifiable Areal Unit Problem

J. K. Nelson[1]

[1]GeoVISTA Center, Department of Geography
Penn State University
335 Walker Building
University Park, PA, 16802
jkn128@psu.edu

## 1. Introduction

The modifiable areal unit problem (MAUP) is statistical bias, resulting from the sensitivity of analytical results of spatial areal data to levels of aggregation (the scale effect), as well as the arbitrary and modifiable sizes, shapes, and arrangements of zones (the zoning effect) (Openshaw 1984a). MAUP was first acknowledged in 1931 (Gehlke and Biehl 1934), yet remains unresolved (Root 2012, Manley 2014). I present an exploratory spatial data analytical (ESDA) approach to understand the scalar effects of MAUP.

Understanding MAUP is crucial to informed analysis of areally aggregated data. Socioeconomic and public health analysts often rely on areally aggregated data, because government regulations on confidentiality prohibit data release at the individual level. Purposeful and meaningful designations of geographic regions for the collection of statistics (i.e. census units) have been under-studied and under-implemented, which arguably degrades the value of census data and the results of analyses using these data (Openshaw 1984b). Inferences about the group are all too often deduced to apply to the individuals that comprise the group aggregate.

The scale at which data are aggregated affects the reliability and validity of conclusions analysts derive from using areally aggregated data (Gregorio et al. 2005). In some cases, understanding relationships between social influences and health may benefit from analyzing both small and large area geographies. There are causal relationships among size, definition, and socioeconomic characteristics of areas and health (Klassen et al. 2004, Root 2012). In selecting the appropriate scale for problem-specific analyses, researchers need to understand the spatial scale at which these causal relationships operate, the stability of operations across scale, and should use multiple scales if there is any uncertainty.

To characterize relationships between data aggregation and spatial scale, I develop a method for statistically and visually exploring the local indicators of spatial association (LISA; Anselin 1995) exhibited between a variable and itself across varying levels of areal aggregation. Such an approach to understanding relationships between MAUP and spatial scale will guide researchers in selecting the most appropriate scales to aggregate and analyze data for problem-specific analyses, and in recognizing the potential for, and reasons that underlie, different analytical answers when analysis is done at different scales.

## 2. Proof-of-Concept

### 2.1 Context

I demonstrate my statistical-visual approach to exploring the scalar effects of MAUP using two proof-of-concept analyses: 2010 median household income in Pennsylvania and 2005-

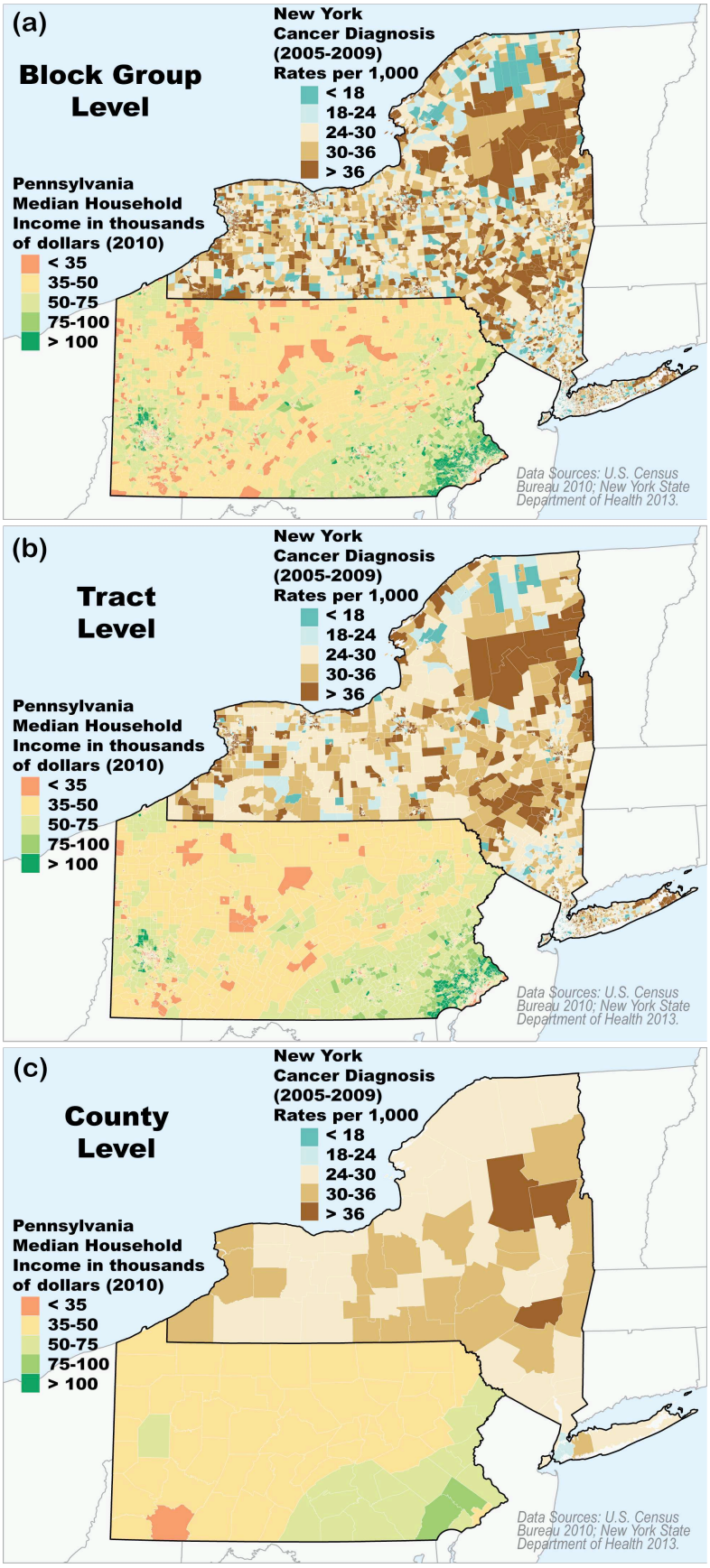



Figure 1: PA Median Income and NY Cancer Diagnosis Rates at the (a) block group level, (b) tract level, and (c) county level.

2009 cancer diagnosis rates in New York. My focus is not to make direct inferences about the spatial phenomena or places chosen for study. Rather, my intent is to provide analysts of areally aggregated data across disciplines with examples of how to quantify, visualize, and interpret the effects of MAUP.

For both analyses, variables' across-scale relationships between county-tract, tract-block group, and county-block group level U.S. geographies are explored. County, tract, and block group levels were chosen for analysis because they cover well-defined, complete state geographies; are statistically more uniform than other census-designated geographies; have more complete and reliable demographic-economic data than other geographies; are commonly analyzed in relevant literature; and foster a more direct assessment of MAUP due to the clean nesting relationship between the three aggregate levels. Figure 1 cartographically depicts both datasets at all three levels. Classification breaks are held constant to visualize the effects of scale on data variability across the three aggregation levels. My aim is to advance the understanding of the (in)stability of these data across the three scales.

## 2.2 Data Processing and Analysis

The first step in achieving this understanding is spatially joining values from upper-level aggregates to nested lower-level aggregates. The spatial joins resulted in: tract-level data with county values appended and block group-level data with county and tract values appended. Next, bivariate LISA analyses were performed. LISA is a local form of spatial autocorrelation, which decomposes the global Moran's I statistic into individual local indicators of spatial autocorrelation. Bivariate LISA analysis allows values of one variable to be regressed on neighboring values of a different variable. Here, I apply LISA to quantify across-scale autocorrelation. As such, the values of the lower-level aggregates were standardized in standard deviation units with a mean of zero and variance of one, and regressed on standardized neighboring values of the appended upper-level aggregate scales.

Spatial autocorrelation analysis quantifies the relatedness of near and distant things, assuming the first law of geography (Tobler 1970). However, defining nearby is challenging and requires an understanding of the spatial phenomena under study. Because my research aims to demonstrate a proof-of-concept rather than make direct inferences about income or cancer, I define neighborhood for both analyses using a first order queen contiguity spatial weights matrix. This is a common approach in the literature; however, designation of both type and order of neighborhood can significantly alter the results of spatial autocorrelation analysis. Thus, the specific outcomes of analysis reported here may differ with alternative definitions of neighborhood.

The ultimate questions the LISA analyses addresses are: how do median household income values or cancer diagnosis rates at one scale correlate with nearby income values or diagnosis rates of a different scale? To explore these questions, statistical output from the bivariate LISA analyses was visually transformed into bivariate choropleth maps (Figures 2 and 3). Special attention to classification breaks and colors was paid to enhance visual exploration of local across-scale (in)stability of median income and cancer diagnosis rates.

The bivariate choropleth maps depict LISA indices against statistical significance, and standardized median income values and cancer diagnosis rates of lower-level aggregates against spatially lagged values and rates of upper-level aggregates. In the maps in column 1 of Figures 2 and 3, purple represents dissimilarity, brown represents marginal spatial autocorrelation, green represents similarity, and lightness in colors signifies level of significance. Darker colors represent greater significance. For example, the dark green areas of southeast Pennsylvania, present in all three across-scale analyses, reflect areas of significant across-scale stability in median income.

In the maps in column 2 of Figures 2 and 3, shades of gray depict similarities between levels of areal aggregation. Outliers, or areas of instability, are depicted in shades of pink and green, where brighter signifies greater disparity. Shades of gray dominate the bottom right maps, illustrating strong similarity in block group-tract relationships for both variables.

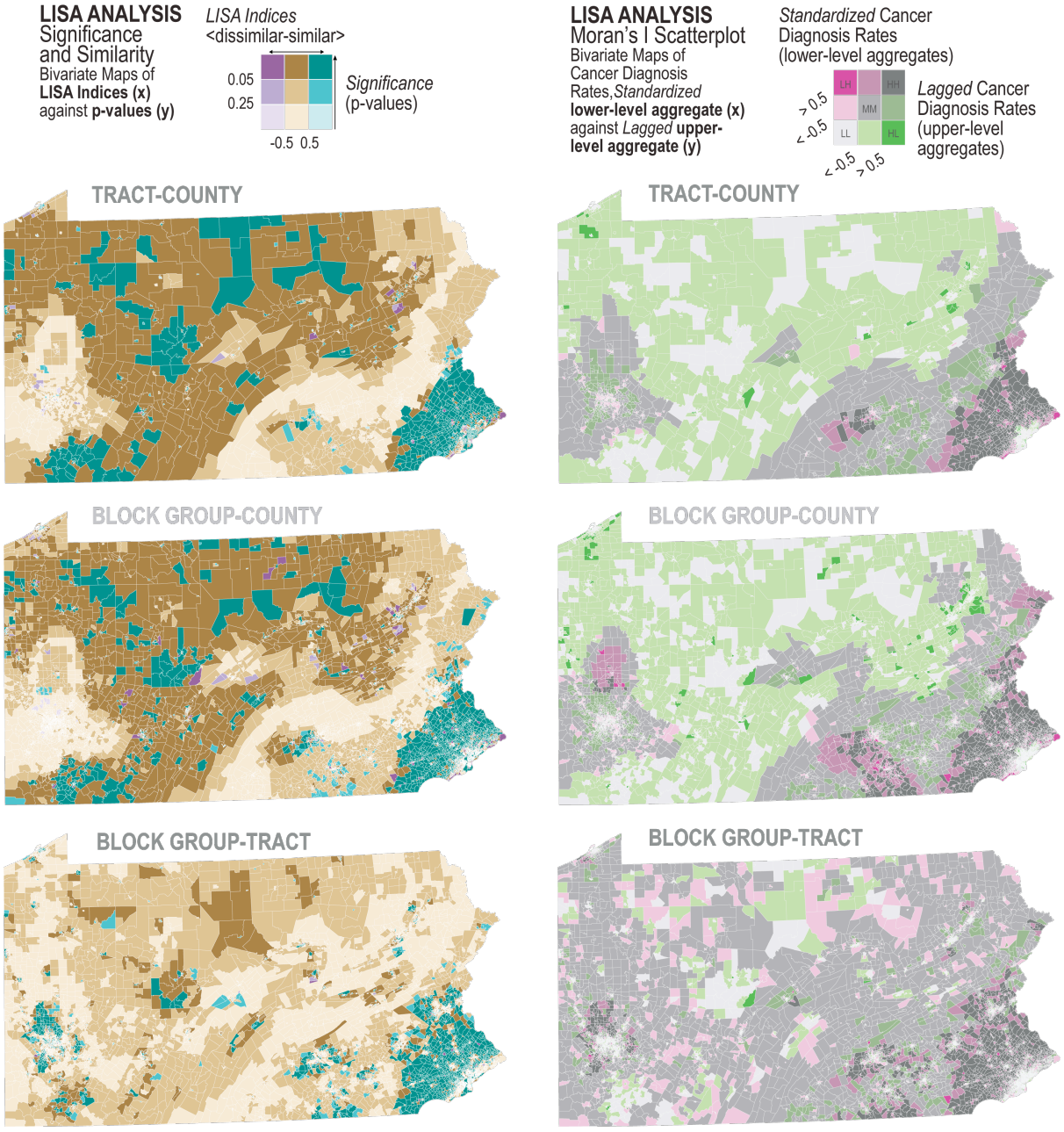



Figure 2: Bivariate choropleth maps of across-scale relationships of PA median income.

## 3. Preliminary Findings

Positive global spatial autocorrelation was found for all across-scale relationships for both studies. The across-scale relationships between tracts and block groups were most similar, because these geographies are more similar in size as compared to counties. The across-scale relationships between counties and block groups were most dissimilar, because these relationships are less direct, and the sizes of their associated geographies most different. Local trends, however, are less predictable and will likely vary by data type. By providing

analysts with a way to visualize local (in)stability they may be able to select appropriate spatial scales for problem-specific analyses and better interpret discrepancies in results when analyzing aggregated data at different spatial scales.

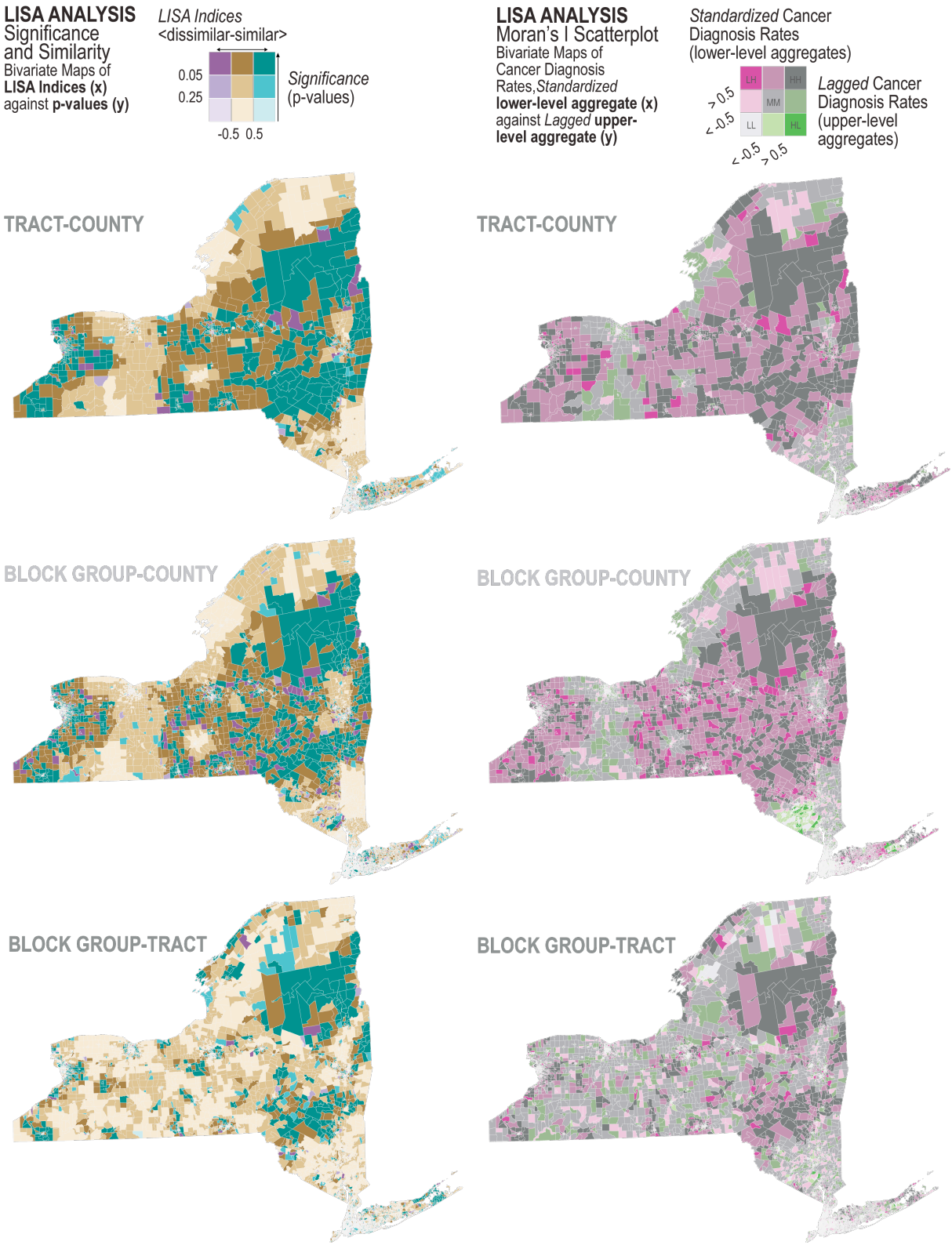



Figure 3: Bivariate choropleth maps of across-scale relationships of NY cancer rates.

In summary, MAUP is a problem going on a century old – an as-yet-unsolved problem – but a very important one to better understand. Characterizing relationships between data aggregation and spatial scale using an exploratory statistical-visual approach can greatly enhance our understanding of MAUP.

## Acknowledgements

This work was supported by the National Science Foundation under IGERT Award #DGE-1144860, Big Data Social Science, and Pennsylvania State University.

## References

Anselin L, 1995, Local indicators of spatial association—LISA. *Geographical analysis,* 27(2):93–115.

Gehlke C and Biehl K, 1934, Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185A):169–170.

Gregorio D, DeChello L, Samociuk H and Kulldorff M, 2005, Lumping or splitting: seeking the preferred areal unit for health geography studies. *International Journal of Health Geographics,* 4(1):6.

Klassen A, Curriero F, Hong J, Williams C, Kulldorff M, Meissner H, Alberg A and Ensminger M, 2004, The role of area-level influences on prostate cancer grade and stage at diagnosis. *Preventive Medicine*, 39(3):441–448.

Manley D, 2014, Scale, Aggregation, and the Modifiable Areal Unit Problem. In *Handbook of Regional Science*, Springer, Berlin, Heidelberg, 1157–1171.

Openshaw S, 1984a, Concepts and techniques in modern geography number 38: the modifiable areal unit problem. *Norwick: Geo Books*.

Openshaw S, 1984b, Ecological fallacies and the analysis of areal census data. *Environment and Planning A,* 16(1):17–31.

Root E, 2012, Moving neighborhoods and health research forward: Using geographic methods to examine the role of spatial scale in neighborhood effects on health. *Annals of the Association of American Geographers,* 102(5):986–995.

Tobler W, 1970, A computer movie simulating urban growth in the Detroit region. *Economic Geography,* 46:234–240.

# Ethnic Structure in Global Naming Networks

K. K. Kowalska[1], P. A. Longley[1], M. Musolesi[2]

[1]Department of Geography, University College London, London, WC1E 6BT, UK
Email: {kira.kowalska.13; p.longley}@ucl.ac.uk

[2]School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK
Email: m.musolesi@cs.bham.ac.uk

## 1. Introduction

In today's multicultural society, there is a growing need to understand the detailed composition of different ethnic groups and the interactions between them. These large-scale group dynamics emerge from the aggregation of millions of ethnic self-identifications of individuals. Mateos et al. (2011) and others have begun to demonstrate the extent to which cultural, ethnic and linguistic (CEL) assignments can be inferred using 'naming networks' of forename-surname pairs of any population.

In further developing this work, we construct a global personal naming network from over 300 million name records from 23 countries. We use this to detect distinct social and ethno-cultural clusters in the network using Louvain community detection algorithm, and examine the interactions between them by inspecting the network structure. The results reveal the degree of isolation, integration or overlap between different human populations, and hence provide new insights into studies on migration, identity, integration or social interaction around the world.

## 2. Methods

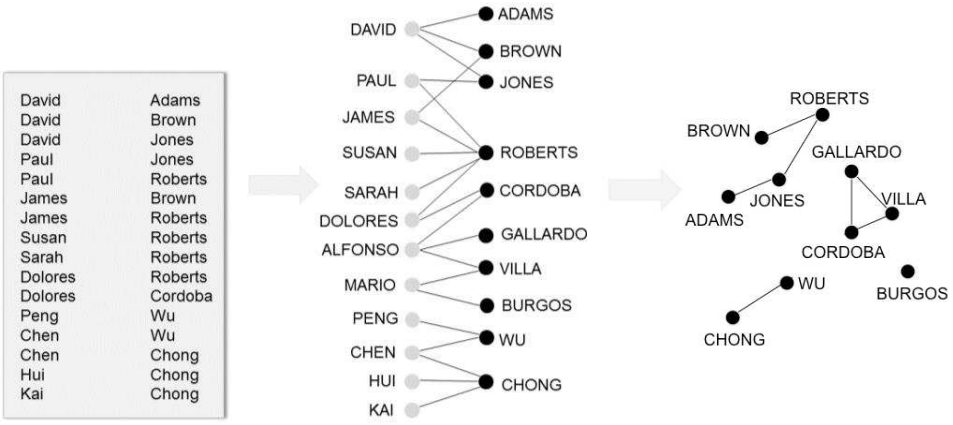
The central rationale to our analysis is that CEL classifications manifest themselves as topological features of networks in which unique surnames are represented as nodes. The subsections below outline how these networks are constructed from raw names data, and how community detection as well as other network statistics can be applied to the 'naming networks' in order to provide insight into the composition and interactions between different ethnic groups.

### 2.1 Names as a network

The first step in our analysis is to visualise names on a network. Given a dataset of people's names, a naming network is formed by treating unique forenames and surnames as network nodes and by placing links between the nodes if a person is identified by a particular forename-surname combination (see Figure 1 (a)).

Having represented names as a network, network links are weighted according to a technique outlined in Mateos et al. (2011). The weighing step adjusts link weights to ensure that very common forenames and surnames do not obscure the network topology, i.e. that the strength of links reflect CEL similarity of names instead of their overall popularity.

Finally, the forename-surname (two mode) network is converted into a one-mode network of surnames only. An example of such a transformation is shown in Figure 1 (b). The weights of the surname network are a result of a simple matrix multiplication of the weights in the forename-surname network (Mateos et al. 2011).




(a) (b)

Figure 1. Converting (a) people's names into a forename-surname network, (b) a forname-surname (two-mode) network to a network of surnames only.

## 2.2 Network community detection

Once names are represented in a surname network format, their ethnic origins are detected using the Louvain network community detection algorithm (Blondel et al. 2008). The algorithm inspects the network structure to unveil clusters of interconnected surname nodes, which can be interpreted as distinct ethnic groups. The Louvain method is chosen for this project because of its ability to deal with very large networks (up to tens of millions of nodes) with weighted links. It is iterative and hence enables investigation of ethnic communities at different levels of resolution. The 'best' level of resolution, i.e. leading to the most distinct communities, is characterized by the highest modularity score (close to 1), where modularity is defined as the fraction of links that fall within the obtained clusters minus the expected value of the fraction if links were distributed at random (Newman and Girvan 2004).

## 2.3 Semi-automated community labelling

Ethnic groups detected using the Louvain method can be automatically assigned their nationality making use of the richness of the naming data used in the project. The data come from 23 countries, enabling the calculation of percentage distribution of each surname across the 23 countries. The average distribution of surnames in one ethnic group could indicate their most probable country of origin. Since not all world countries are present in the data, some statistics are needed to decide whether a community comes from one of the represented countries, making automated labelling possible, or not, hence leaving the labelling to human expertise. The statistics investigated in the project, based on average percentage distributions of communities (or surnames they contain) across countries, are:

1. Percentage in dominant country (Geographic Dominance)
2. Standard deviation across countries (Geographic Spread)
3. Mean cosine similarity of surnames assigned (Geographic Integrity)

The assumption is that communities with high geographic dominance, high geographic integrity and low geographic spread could be automatically assigned nationality of their dominant country. The remaining communities could represent nationalities missing from the data or

ethnic groups that do not belong to any country (e.g. Romani people), and hence would require further investigation.

### 2.4 Network measures of interaction

Node properties of degree, betweenness and farness (Newman, 2010) are used to quantify interactions between surnames (nodes) in the naming networks. Surnames with high degree share similar naming practices with a large pool of other surnames, and hence might belong to a large ethnic group or one with unusually large surname heterogeneity. Surnames with high betweenness play an important role in connecting different parts of the global naming network and hence could be called 'cultural connectors' as they are most likely to interact with people from different ethnic backgrounds. Finally, surnames with high farness are least integrated within the global community; average farness of a community could be used as a measure of CEL isolation.

### 2.5 Data

Data used in this analysis come from a very extensive database of over 300 million people's names from 23 countries in four continents, collected from telephone directories and electoral registers and analysed as part of the 'Uncertainty of Identity' project at University College London (http://www.uncertaintyofidentity.com/). The data represent each country at a varying level of accuracy (i.e. the percentage of total population captured for various countries ranges from 0.3% to 79%). Therefore, before constructing the global naming network, name frequencies from each country are proportionally weighted to represent their total country's population.

## 3. Results

The world names data were converted into a naming network according to the steps outlined in Section 2.1. Firstly, a two-mode network was created with unique forenames and surnames as nodes (1,497,327 forename, 1,128,970 surname nodes). The two-mode network was then converted to a one-mode network of surnames only, which was subsequently used for the analysis of ethnic population structure around the world.

Ethnic communities in the global naming network were detected using the Louvain method. The algorithm started by assigning each surname node to a separate community, and then iteratively merged highly-connected communities until it arrived at the maximum modularity score of 0.628303. The resulting partition consisted of 7,947 ethnic communities of sizes varying from 2 to 157,889 surnames.

The geographic properties of dominance, spread and integrity of the ethnic communities were quantified using the statistics introduced in Section 2.3 (see Figure 2) in order to select communities suitable for automated nationality labelling. Varying thresholds on the statistics had impact on the number of surnames labelled, as shown in Figure 3.

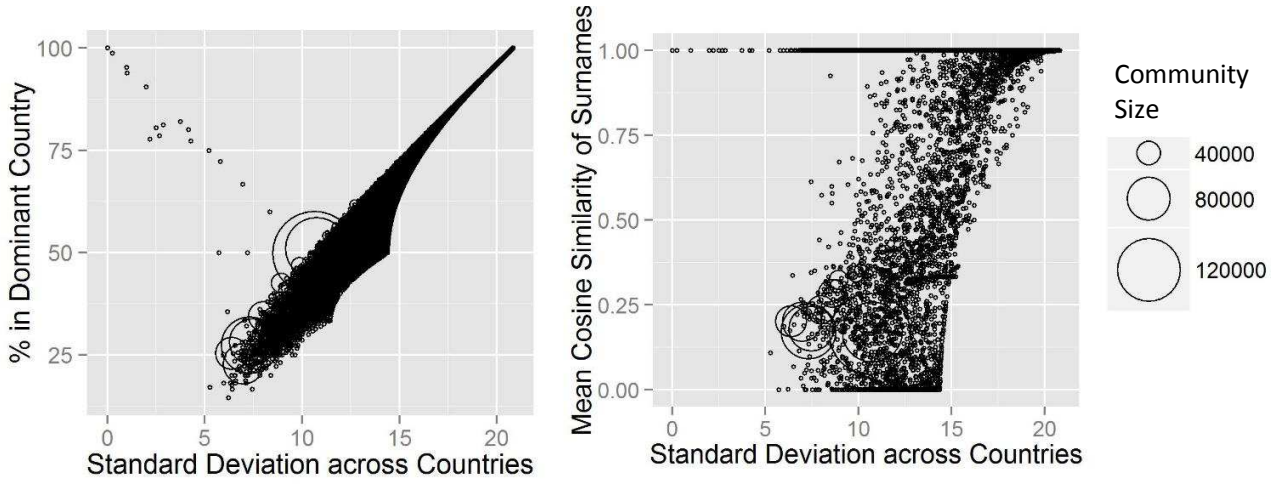



Figure 2. Ethnic communities scattered according to their geographic dominance (% in dominant country), spread (standard deviation) and integrity (mean cosine similarity of their surnames).

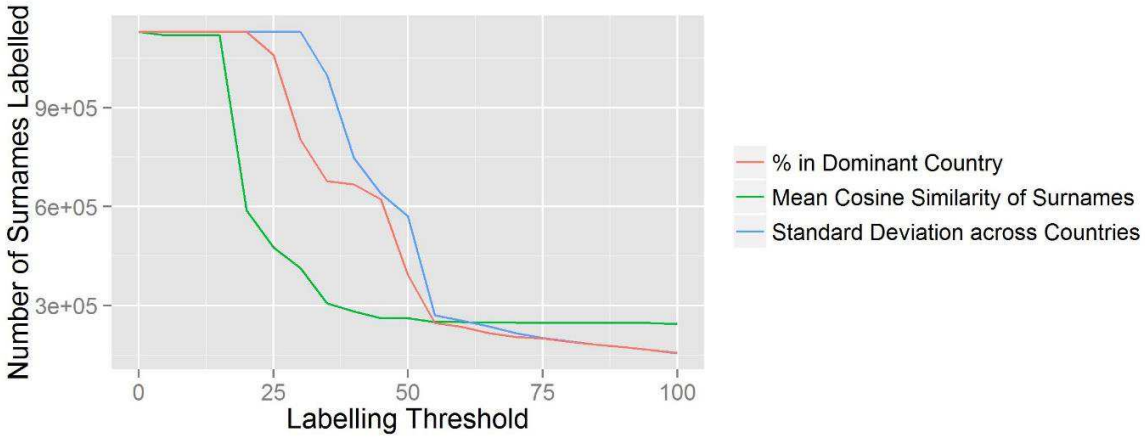



Figure 3. Number of surnames automatically labelled for different % thresholds on the three statistics of ethnic communities.

Properties of individual surnames were investigated by measuring their degree, betweenness and farness. Surnames corresponding to the most extreme values of the three statistics are summarized in Table 1. As discussed in Section 2.3, surnames ‘Le’, ‘Begum’ could be labeled as cultural connectors, whereas surnames ‘Markovic’ are ‘Jankovic’ represent most culturally or linguistically isolated individuals in the retained data.

Table 1. Surnames with extreme network properties.

| Node Property | Highest | Lowest |
|---|---|---|
| Degree | Patel, Khan | Rahmani, Minar |
| Betweenness | Le, Begum | Laib, Bouaka |
| Farness | Markovic, Jankovic | Patel, Begum |

## 4. Conclusions and Future Work

The paper presents preliminary results of analysing topology of 'naming networks' in order to gain insight into ethnic population structure around the world. The work is still in progress and numerous future directions are possible. In the first instance, validation techniques are needed for the automated labelling presented in this paper.

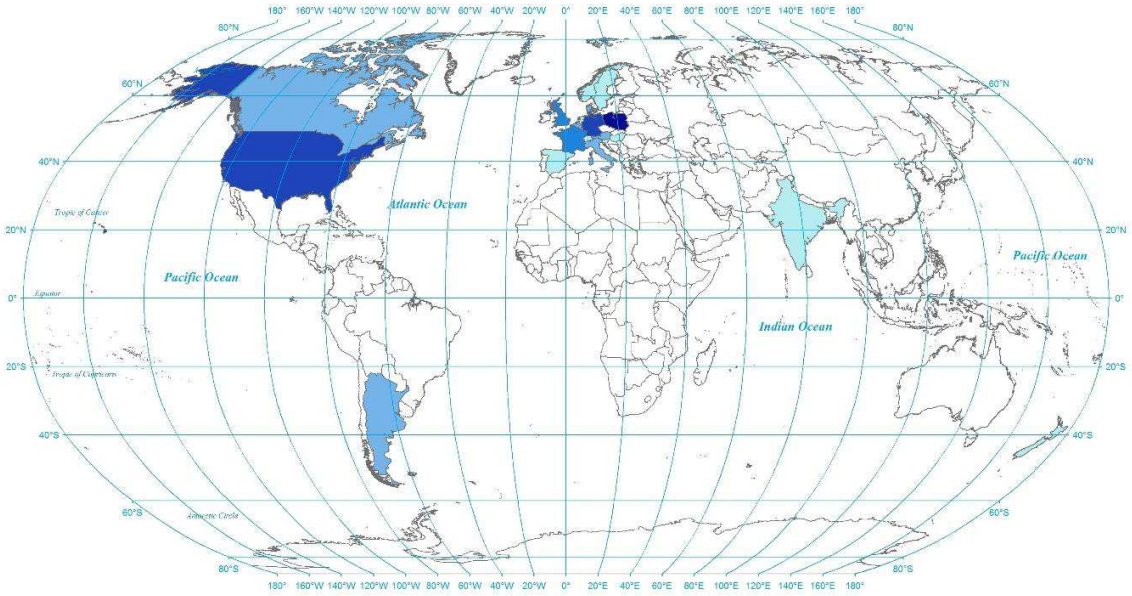



Figure 4. Concentration of surnames classified as Polish when all communities with >40% in their dominant countries are automatically labelled (uncoloured countries are not present in the world names data).

## References

Blondel VD, Guillaume J, Lambiotte R and Lefebvre E, 2008, Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008.

Clauset A, Newman MEJ and Moore C, 2004, Finding community structure in very large networks. Physical Review E70, 066111.

Mateos P, Longley PA and O'Sullivan D, 2011, Ethnicity and population structure in personal naming networks. PLoS One, 6(9): e22943.

Newman MEJ, 2010, Networks: An Introduction. Oxford University Press, Oxford, UK.

Newman MEJ and Girvan M, 2004, Finding and evaluating community structure in networks. Physical Review E69, 026113.

# How Representative are Social Media Datasets of the True Population: A Case for London

A. B. Leak[1], M. Adnan[1], P. A. Longley[1]

[1]University College London, Gower St. London, WC1E 6BT
Email: {a.leak.11;m.adnan;p.longley}@ucl.ac.uk

## 1. Introduction

The study of personal names offers valuable insight into the structure and dynamic of human populations (Longley et al. 2011). The reason being that personal names are a strong indicator of their bearer's cultural, ethnic and linguistic characteristics (Mateos et al. 2011).

The utility of this information inherent within personal names has led to the development of the World Names Database (WND: http://worldnames.publicprofiler.org/); a synthesis of National Census and Telephone Directory data for twenty-six countries across four continents accounting for approximately 2 billion of the Earth's population. Whilst offering a significant step forward in the analysis and understanding of global populations, the WND is limited by its current inclusion of countries. The primary challenge in this respect being the availability of publicly available registers.

This study forms part of an overarching goal to develop proxy population registers by using publically available online social media data. In this instance, the platform being demonstrated is Twitter. Launched in 2006, Twitter has 255 million active users of which 78% accesses the service through a mobile device (Twitter, Inc. 2014).

This paper will briefly outline the creation of such a population register for Greater London at Borough and Parliamentary Region levels, the methods used to assess its representative capability, and finally present preliminary results of this analysis with a discussion as to its potential strengths and weaknesses. Proving successful, the methodology proposed will be applied to further countries, as yet not included in the WND.

## 2. Methodology

### 2.1 Data Processing

Creation of the new population register falls into two main phases. First, the extraction of a viable forename and surname from a user's display name and second, the identification of their most likely residential location. Based on the work of Adnan et al. (2013), the name extraction methodology applies an intelligent heuristic algorithm to clean and segment a user's name into their individual components. This method is centred on the principle of western naming order – forename followed by surname – and takes into account many cultural naming conventions. For example, the surname prefixes De, De la, and Van. Table 1 presents a sample of Twitter names and their extracted components.

Table 1. Sample of forename-surname pairs extracted from usernames

| Twitter Name | Given Name | Surname | Real Name |
|---|---|---|---|
| Rhiannon De la Merr | Rhiannon | De la Merr | Yes |
| Mystic Meg | Mystic | Meg | No |
| James Evans 1989 | James | Evans | Yes |

Probable residential location is assigned to each user as the administrative unit in which the user broadcast the majority of their Tweets and the number of Tweets within the administrative unit is greater than or equal to 5 and where greater than 50% of the users total Tweets fall within the same administrative unit. This method restricts each user to only one possible location, whilst minimizing the number of tourists and non-residents captured within the analysis. The Twitter dataset used in the study was collected as part of the Uncertainty of Identity project at UCL between September 2012 and January 2014 through the Twitter API Sample Stream. It should be noted that whilst the sample returns only 1% of the total service throughput, where only geo-referenced Tweets are requested, this accounts for roughly 1% of all Tweets and as such a high proportion of all spatial Tweets are returned (Morstatter et al. 2013).

### 2.3 Similarity Measurement

The representative capability of the Twitter based population register will be assessed at multiple spatial resolutions using the District and Parliamentary Region administrative boundaries against the 2011 Enhanced Electoral Roll (EER) provided by CACI Ltd (http://www.caci.co.uk/). The EER contains forename, surname and address information for the majority of the UK population eligible to vote. In each case, similarity between the Twitter and Electoral Roll population for each administrative unit will be calculated using the Morisita-Horn index of overlap. The Morisita-Horn method, brought from ecology, is recognised for being unbiased when dealing with differing sample sizes (Wolda 1981). Equation 1, the Morisita-Horn index returns a linear index where 0 indicates no overlap in surname composition whilst 1 indicates an identical composition of surnames.

$$C_H = \frac{2\sum_{i=1}^{S} x_i y_i}{\left(\frac{\sum_{i=1}^{S} x_i^2}{X^2} + \frac{\sum_{i=1}^{S} y_i^2}{Y^2}\right) XY}. \tag{1}$$

**S** is the number of unique surnames shared between the two populations
**xi** and **yi** are the number of individuals sharing a specific surname in region x and y
X and Y are the number of unique surnames in regions X and Y respectively

## 3. Results and Discussions

The results from this investigation clearly illustrate the limitations inherent with such a data source as a means to reconstruct national population registers. Figure 2, the map of location quotient illustrates where Twitter under and over represents the true population. Particularly evident is the City of London, which has a Twitter population four times greater than expected. This feature is indicative of an area offering significant employment yet minimal residence. Further, the Outer East and North East areas of London appear to perform more poorly. Figure 3 shows a moderate positive relationship of r=0.50 between Twitter and EER populations at Borough level however this falls significantly to r=0.16 when compared for Parliamentary Region. The implication of this being that the ability for a social media dataset to infer population density is feasible however only at a limited spatial resolution.

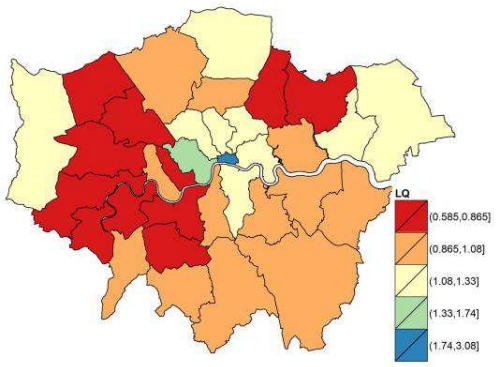
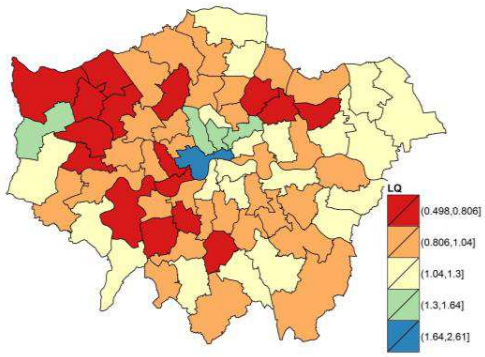
| Borough | Parliamentary Region |
|---|---|
|  |  |



Figure 2: Location quotient for Twitter vs Electoral Roll population at Borough and Parliamentary Region levels.

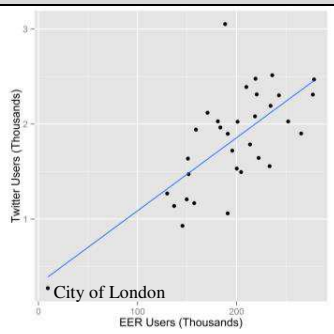
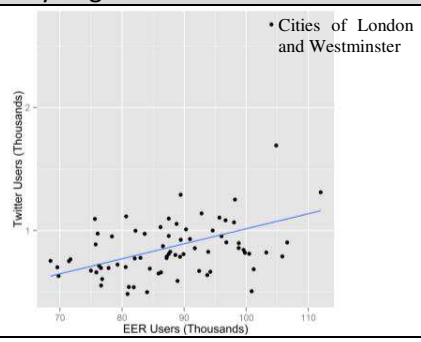
| Borough | Parliamentary Region |
|---|---|
|  |  |
| R-squared: 0.5018*** | R-squared: 0.1646*** |



Figure 3: Scatterplot of Twitter vs Electoral Roll population counts at Borough and Parliamentary Regions levels.

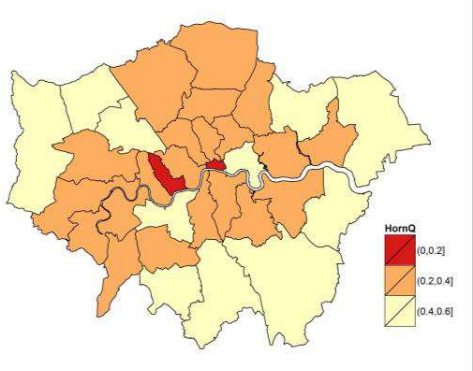
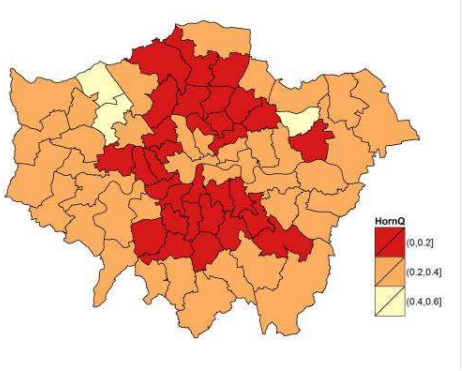
| **Scale** | | Borough | Parliamentary Region |
|---|---|---|---|
| **Map** | |  |  |
| **Similarity** | **Min** | 0.15 | 0.10 |
| | **Mean** | 0.36 | 0.24 |
| | **Max** | 0.52 | 0.60 |



Figure 4: Morisita-Horn similarity between Twitter derived and Electoral Roll population registers at District and Parliamentary Region levels.

Figure 4 clearly demonstrates the rapid decrease in mean similarity as the spatial resolution is increased. In fact, only when we consider Greater London as a distinct unit do we approach a reasonable level of accuracy with a similarity score of 0.69.

Based on the evidence that areas with a larger resident population perform more favourably, it may be postulated that through the amalgamation of administrative units, which perform poorly, the mean level of similarity could be increased to an acceptable level. Whilst this may not be feasible where no existing data are available for verification, it may prove useful in creating custom geographical regions for analysis purposes.

The decrease in similarity at higher spatial resolutions was reflective of a previous study based on 2007 Electoral Roll data by Leak et al. (2014) which saw similar degradation of similarity as the spatial resolution was increased. Whilst this previous study used EER data from 2007, the District level results for the Greater London were largely similar. The previous study also observed that Greater London area was one of the poorest performing nationally.

## 4. Conclusions

This paper has presented a provisional analysis of the representative capability of a social media based population register within Greater London against the 2011 EER at two administrative geographies. It is fairly evident from this preliminary analysis that the Twitter based population register struggles at higher spatial resolutions as the limited pool of users is increasingly segmented. This being said, the objective of the study was to create population registers for countries where data is currently unavailable through conventional means. With this in mind, the analysis has proven that, at an appropriate spatial resolution, online social media may offer a viable solution.

With an improved understanding of the nuances of social media based population registers, further steps can be taken toward filling in countries not yet in the World Names Database. This may in turn allow for the analysis of migration patterns between countries and offer new insights into population dynamics where traditional population registers are not yet publically available.

## Acknowledgements

This work was completed as part of the EPSRC research grant "The Uncertainty of Identity: Linking Spatiotemporal Information in the Real and Virtual Worlds" (EP/J005266/1) and the DSTL PhD scheme (12/13NatPHD_61).

## References and Citations

Adnan, M., Lansley, G., Longley, P.A., 2013. A geodemographic analysis of the ethnicity and identity of Twitter users in Greater London.

Leak, A., Adnan, M., Longley, P., 2014. Towards a Seamless World Names Database, in: Demography, Identity and Dynamics 2. Presented at the AAG Annual Conference.

Longley, P. A., Cheshire, J. A., & Mateos, P. (2011). Creating a regional geography of Britain through the spatial analysis of surnames. Geoforum, 42(4), 506-516.

Mateos, P., Longley, P.A., O'Sullivan, D., 2011. Ethnicity and Population Structure in Personal Naming Networks. PLoS ONE 6, e22943. doi:10.1371/journal.pone.0022943

Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M., 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. Proc. ICWSM.

Twitter, Inc., 2014. Twitter Reports First Quarter 2014 Results.

Wolda, H., 1981. Similarity indices, sample size and diversity. Oecologia 50, 296–302.

# Crowdsourcing Landscape Perceptions to Validate Land Cover Classifications

Kevin Sparks[1], Alexander Klippel[1], Jan Oliver Wallgrün[1], David Mark[2]

[1]The Pennsylvania State University, University Park, State College, PA 16801
Email: {kas5822; klippel; wallgrun} @psu.edu

[2]NCGIA & Department of Geography, University at Buffalo, Buffalo, NY 14228
Email: dmark@buffalo.edu

## 1. Introduction

This paper analyzes the correspondence between human conceptualizations of landscapes and spectrally derived land cover classifications. Although widely used, global land cover data have known discontinuities in accuracy across different datasets. Computational accuracy assessments are performed to correct for errors, yet inaccuracies and disagreement persist (Foody 2002). With the emergence of crowdsourcing platforms large-scale contributions to validate land cover classification are now possible and practical. The potential use of crowdsourcing methods for validation purposes by having human volunteers check for inconsistencies in global land cover datasets has been recognized by previous research. The Geo-Wiki project (Fritz 2009), for instance, asks online participants to use aerial imagery via Google Earth as well as their local knowledge to validate whether or not the land cover/land use is being accurately represented by the land cover classification in question. This volunteer geographer approach complements the accuracy assessments in use, but fails to guarantee a level of quality in the volunteered data. If crowdsourced human participants are to be incorporated into accuracy assessments of land cover types, there needs to be some understanding of how humans perceive and conceptualize land cover types and rigorously measure how well humans perform in recognizing predefined land cover classifications.

We are reporting on three experiments that provide insights on the relationships between human conceptualizations of landscapes and land cover classifications using novices, educated novices, and experts. Our findings suggest misclassifications are not random but rather systematic to unique landscape stimuli and unique land cover classes. By comparing novices and experts we are able to evaluate the potential for using crowdsourcing in aiding land cover classifications.

## 2. Methods

Two datasets were used for this experiment: on-the-ground-photographs of landscapes provided by The Degree Confluence Project (DCP) (confluence.org), and the National Land Cover Dataset (NLCD) 2006 (Fry 2011) provided by the Multi-Resolution Land Characteristics (MRLC) consortium.

The DCP is a site that provides a platform for collecting crowdsourced photographs of landscapes at confluence points across the world in a systematic way. Research methods have shown a successful level of reliability in using DCP data for validating land cover classifications (Iwao 2006). Confluence as defined for the purposes of the DCP is the location where two integer latitude and longitude coordinates meet. An example of this would be ‘latitude 42 N, Longitude 100 W’ as opposed to ‘latitude 42.65 N, longitude 100.23 W’. Users are encouraged to visit these

locations, take photographs of the landscape, and upload the images with metadata such as date visited.

We constrained our data collection to the continental United States. A total of 799 photographs where collected out of a possible 856. Two sampling criteria restricted the data collection process: First, scenes that included snow in the photograph were excluded as this is not reflective of the landscape or land cover but rather temporal weather conditions. Second, images that included human presence were excluded. Outside of these sampling restrictions, few confluences do not have photographs uploaded to the website, and as such, could not be collected.

Latitude and longitude coordinates from the DCP dataset were extracted and converted into a point shapefile to be used in ESRI's ArcGIS software. This allowed for the extraction of the corresponding land cover class from NLCD level II (16 land cover classes) for each confluence point and the corresponding image. Although land cover change has the possibility of influencing incorrect land cover extraction, each of the 77 images were visually analyzed with their corresponding land cover class to ensure logical consistency. It is important to note that Wickham's (2013) accuracy assessment of NLCD 2006 for the conterminous United States concludes that level II accuracy = 78%. For the scope of this experiment we aggregated Deciduous-Forest, Evergreen-Forest, and Mixed-Forest into one "Forest" class, leaving a possible 14 land cover classes to sample from.

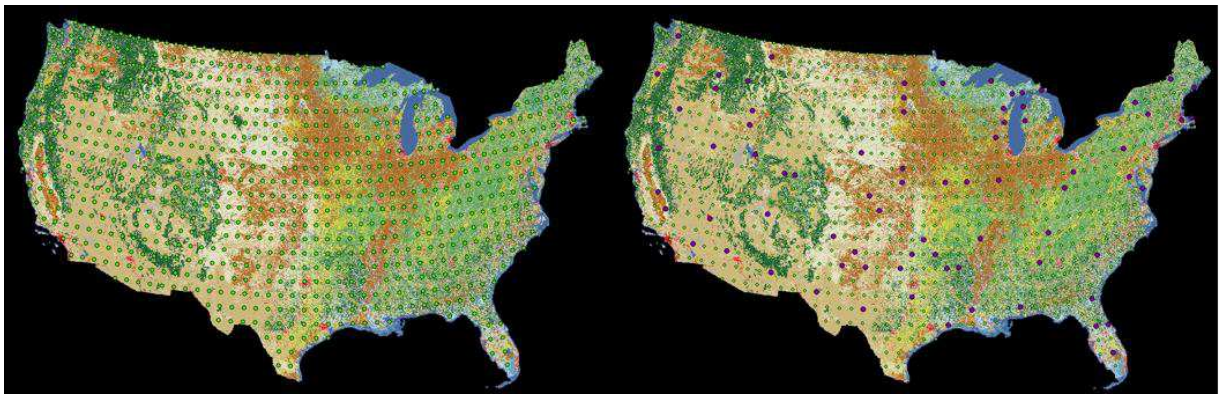


Figure 1. The NLCD 2006 overlaid by confluence points (left). Stratified random sampled confluence points, 77 total sampled, 7 in each land cover class (right).

The images, now each defined with a land cover class, were sorted into bins based on their land cover class. A total of 11 land cover classes were sampled from a possible 14 with Developed, Medium-Intensity, Developed High-Intensity, and Perennial Ice/Snow not represented. Seven images from each class were randomly selected, totaling 77 images.

The experimental software CatScan (Klippel 2008) used for the experiment has been designed to be serviceable in combination with Amazon's Mechanical Turk (AMT). In each experiment, participants performed a non-free classification task. All images were initially displayed on the left panel of the screen. On the right side of the screen, the 11 land cover classes were displayed into which participants were able to drag icons from the left panel. It was possible to leave classes empty.

Three experiments were conducted to provide insight on the relationships between human perceptions/conceptualizations of landscapes and land cover classifications. The first experiment solicited 20 lay participants (5 female) to perform the non-free classification test. The second

experiment (20 lay participants, 11 female) included an intervention of definitions and visual examples for each land cover class. The third experiment used expert participants only (4 experts, ecological and GIS backgrounds with experience in working with land cover).

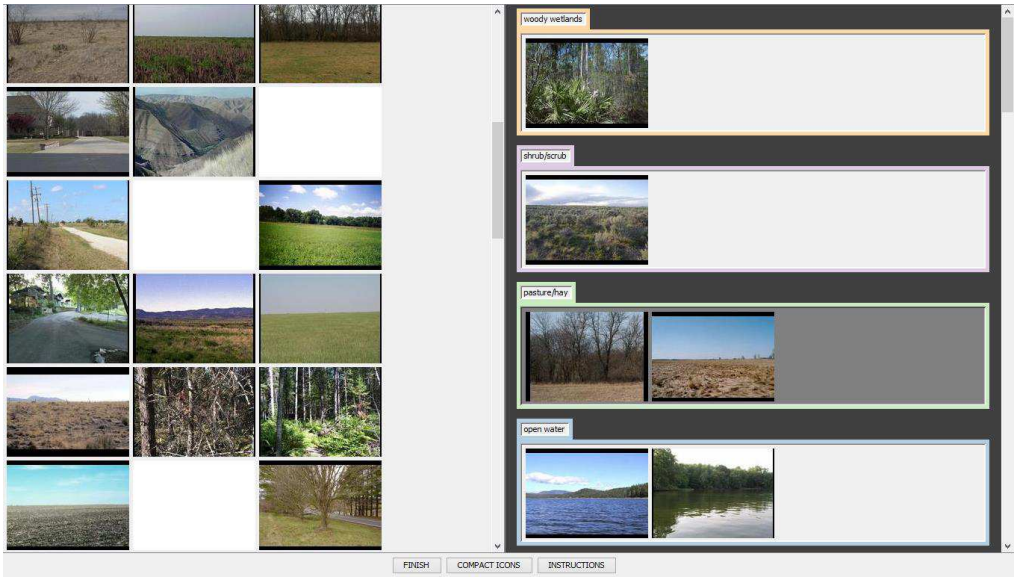



Figure 2. Screenshot of the CatScan interface of an ongoing mock-up experiment.

## 3. Results

To analyze the classification results, we created confusion matrices (Figure 3-5) that show the number of correctly classified land cover images and in which classes the misclassifications occur. We performed chi square tests to corroborate the interpretation statistically. Several main observations can be summarized as follows: First, there are statistically significant differences between the number of 'correctly' (the NLCD class is considered 'correct') identified land cover images across all 11 land cover classes in all three experiments. Second, the improvement in classification of lay participants after the intervention is statistically significant ($\chi^2 = 5.2807$, df = 1, p = .02). Third, there is no statistically significant difference between educated lay participants and experts ($\chi^2 = 1.52$, df = 1, p = .22). Forth, the overall match between participants' classifications and NLCD is rather low (40.19 - 48.37%). This accuracy rate range still indicates the difficulty of human land cover classification even in the face of measuring it against the inaccuracies of NLCD 2006. If human classification was near perfect, then we would expect to see accuracy rates of approximately 78%, matching the NLCD 2006 level II accuracy rate.

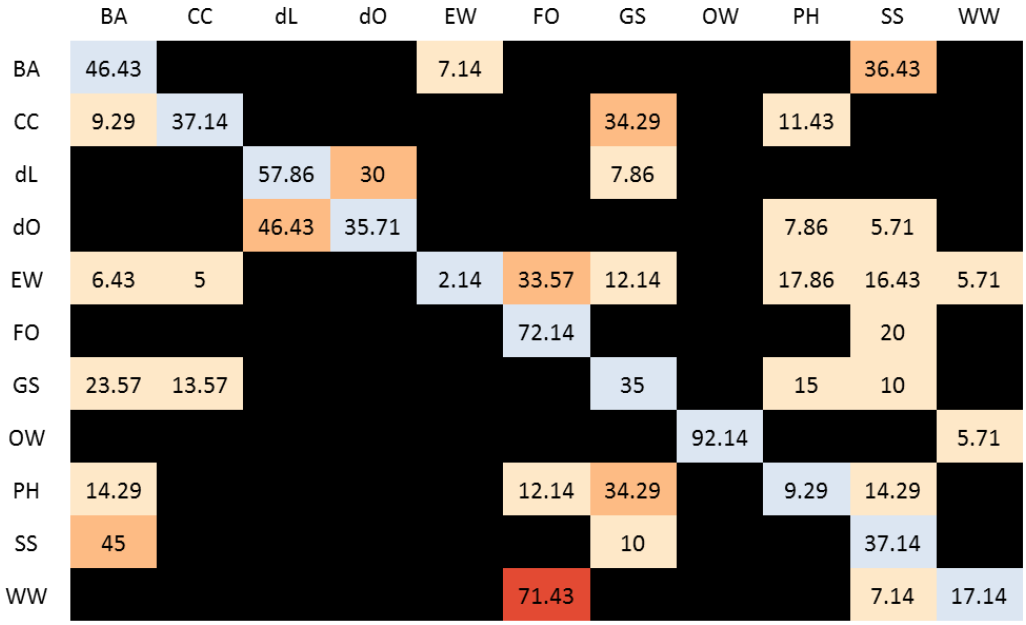



Figure 3. Confusion matrix for experiment 1 showing percentages of correct (diagonal) and misclassified landscape images (rows). Total number of misclassified images smaller/equal 5% are blackened out, values between 5% and 25% are indicated by light pink areas; values between 25% and 50% are light orange and, misclassifications above 50% are red.

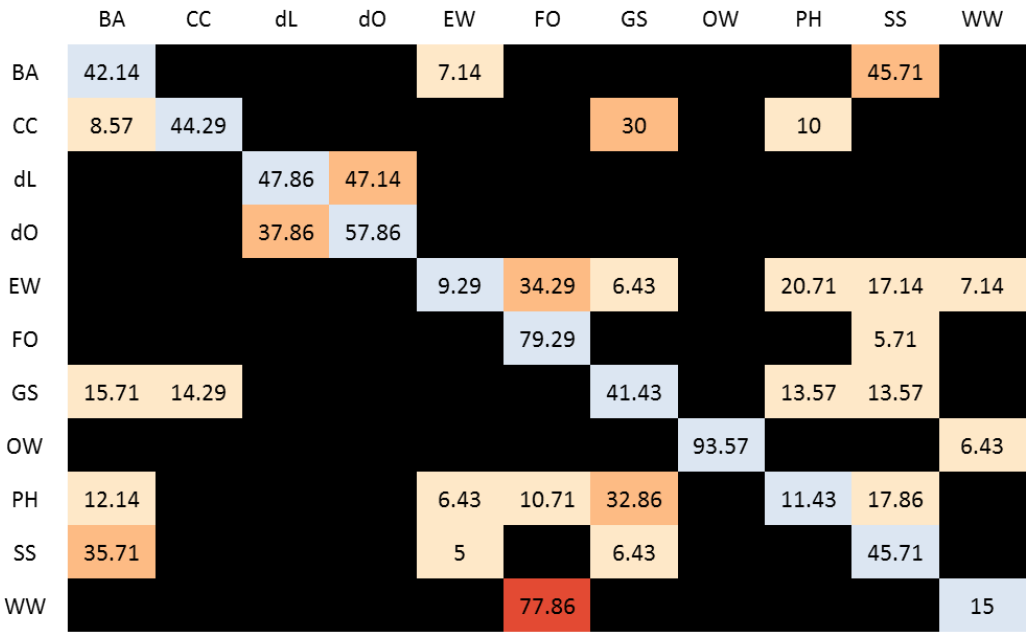



Figure 4. Confusion matrix for experiment 2.

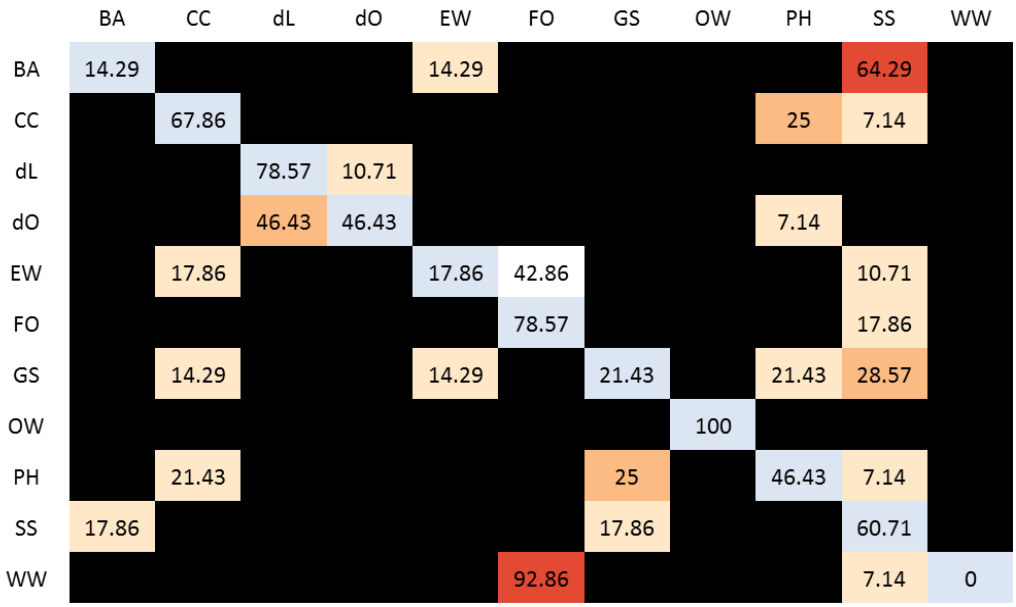



Figure 5. Confusion matrix for experiment 3.

## 4. Discussion / Outlook

The overall accuracy increased statistically significantly using an intervention of providing definitions and prototypical images as examples as mentioned previously. The misclassifications are not random but rather systematic. This is the case on the level of land cover classes as well as on the level of individual images.

When assigning complex tasks to be performed by the crowd, one must ensure that the volunteered data quality is appropriate and sustainable. In the context of land cover validation, humans are very successful in correctly classifying certain landscapes via on the ground photographs, and poor in classifying others. Lessons learned from these three experiments are currently integrated in additional experiments that will, among other things, provide additional information about the area to be classified in form of aerial images, ask participants to perform classifications along individual dimensions, and allow for an indication of uncertainty of classifications.

## Acknowledgments

This research is funded by the National Science Foundation under grant #0924534

## References

Foody G M, 2002, Status of land cover classification accuracy assessment. Remote Sensing of Environment, 80(1):185-201.

Fritz S, Mccallum I, Schill C, Perger C, Grillmayer R, Achard F, Kraxner F, and Obersteiner M, 2009, Geo-wiki.org: The use of crowdsourcing to improve global land cover. Remote Sensing, 1(3):345-354.

Fry J, Xian G, Jin S, Dewitz J, Homer C, Yang L, Barnes C, Herold N, and Wickham J, 2011, Completion of the 2006 national land cover database for the conterminous United States. PE&RS, 77(9):858-864.

Iwao K, Nishida K, Kinoshita T, and Yamagata Y, 2006, Validating land cover maps with degree confluence project information. Geophysical Research Letters, 33(23).

Klippel A, Worboys M, and Duckham M, 2008, Identifying factors of geographic event conceptualisation. International Journal of Geographic Information Science, 22:2, 183-204

Wickham J D, Stehman S V, Gass L, Dewitz J, Fry J A, Wade T G, 2013, Accuracy assessment of NLCD 2006 land cover and impervious surface. Remote Sensing of Environment, 130, pp. 294–304.

# Modelling the Fuzzy Footprints of Villages from Postal Address Records

F.M. Almadani[1], P.F. Fisher[1], and C.H. Jarvis[1]

[1]Department of Geography, University of Leicester
Leicester LE2 3TD
Email: {Fma7, Pff1, Chj2}@Leicester.ac.uk

## 1. Introduction

Historically British villages would have been defined by the parish boundaries and hence would have been bounded objects in some sense. Empirical data like the Ordnance Survey AddressPoint data, includes village names which are listed as a part of the address information. In some cases the villages so named are not identified by the parishes while other villages spill over into the area defined by another parish; they are geographically vague, both in definition and extent, as they often do not have officially defined boundaries or have formal names. Here, we develop modelling techniques that adequately handle vagueness and imprecision in the naming of villages. Particularly we propose two different approaches to approximate the fuzzy spatial extent (fuzzy footprint) of villages.

## 2. Prior Work on Modelling Vague Regions

Crawford (2002) developed a novel approach for linking population and environmental data across a thematic domain, and representing functional regions in the landscape as village territories. Similarly, Gray (2008) generated fuzzy models of a settlement's spatial extent, using settlements' place names gained from address point data for postcodes. In the same vein, Chaudhry and Mackaness (2008) presents a method for automatically identifying settlement boundaries based on what typically constitutes 'citiness'.

Moreover, there is much work in the fuzzy literature to identify vague regions using qualitative methods engaging human subjects (Montello et al. 2003; Lüscher and Weibel 2013). Furthermore, a number of studies use information from the web to model the extent of vague places (Goodchild et al. 1998, Hollenstein and Purves 2010). Most of these studies are based on methods that provide a density surface as a representation of the vague region. Here, we pursue a density surface modelling approach in a different context and in greater depth (see Section 4.1 onwards); we model human settlements derived from postal addresses rather than from the web (tag points) or human subjects (questionnaires).

## 3. Data specification

The study area of this research covers the rural settlements of Hinckley and Bosworth District in the southwest area of Leicestershire in England. Address point data for this area were obtained from the Ordnance Survey (OS MasterMap ® Address Layer 2). These data consist of points for every postal address for every house (See Figure 1 for some examples). Parish boundaries were also used to act as the official boundaries for rural settlements. According to a definition provided by the Department for Community and Local Government (2010) a parish name

refers.to the geographical name of an area which might be known locally as a town, community, neighbourhood or village

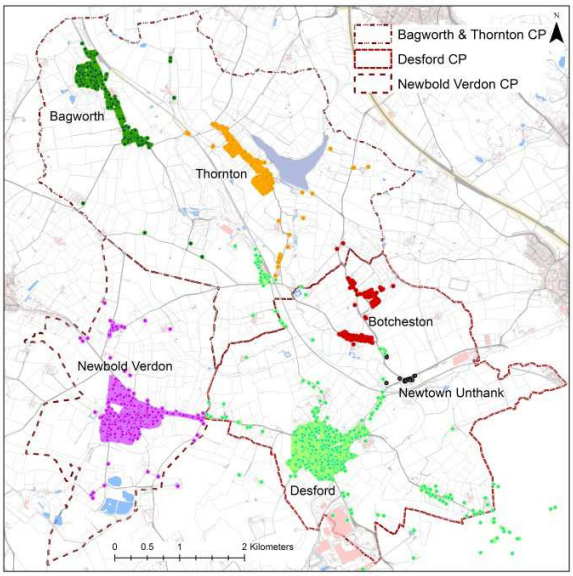



Figure 1: Maps of extracted villages form the Address Point records with some equivalents Parish boundaries.

## 4. Modelling the Fuzzy Footprints

### 4.1 Modelling based on Density of Houses

A common approach applied in GIScience is Kernel Density Estimation (KDE) (Hollenstein and Purves 2010,). KDE basically generates a smooth, continuous surface from point patterns that represents the spatial variations of events (Silverman 1986). It is used here to model the spatial footprint of the "fuzzy extent" of villages by estimating the proportion of total incidents (addresses) that can be expected to occur at any given map location.

Essentially the choice of the surface resolution and the bandwidth in this function both influence the shape of the surface in terms of its smoothness or peakness (Silverman 1986). The KDE for each village was computed independently using the standard distances as the bandwidth. Figure 2 illustrates the application of KDE to some rural settlements.

To represent the spatial extent of villages as fuzzy objects, a further step is needed to transform the density surface to a fuzzy-set based approach (Zadeh 1965). The fuzzy memberships from the density surface are found by normalisation; this scales the range of the KDE results to 0 to 1 (Figure 4 D-F).

### 4.2 Modelling based on distance

An alternative is inspired by fuzzy classification; each object is spread out over the various clusters by means of a degree of belonging that is quantified by membership coefficients that range from 0 to 1 (Kaufman and Rousseeuw 1990). As we are assuming that a set of address points sharing a village name represent one single cluster, we can consider the centroid as the core of that village. So from the address points in a village the mean centre is identified, and for every house the distance away from the village centre is calculated. Figure 3 provides some mapping results of these distances.

.

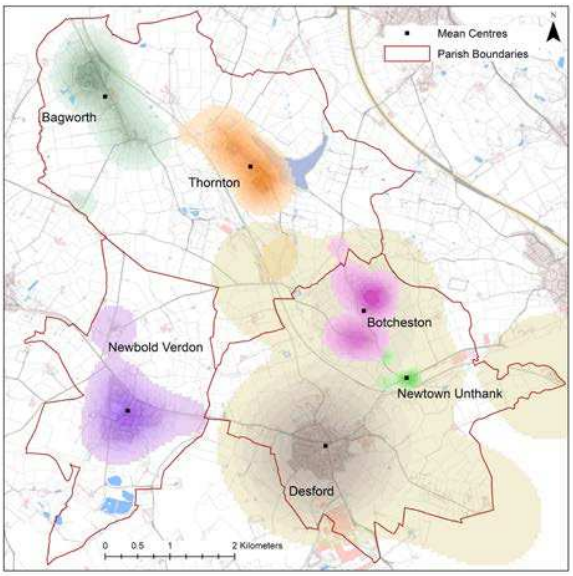



Figure 2: Maps of Address Point data for some settlements and their fuzzy footprints resulted from KDE modelling.

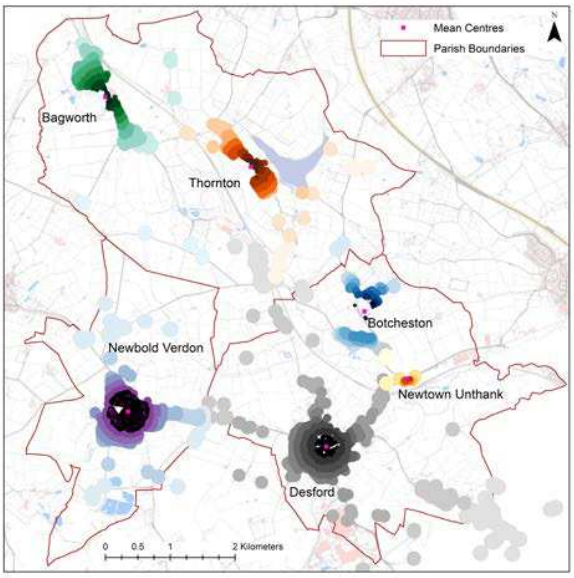



Figure 3: Maps of Address Point data for some settlements based on the distance values from their mean centres.

Interpolation from the address points extends the memberships to a continuous surface over the entire area within that village. Following Hall et al. 2011, ordinary kriging is used to transform these point measurements into the continuous field representation. Again normalisation is necessary to transform values to fuzzy memberships. The membership values vary according to the distance of the location from the village mean centre, as can be seen in Figure 4(G-L).

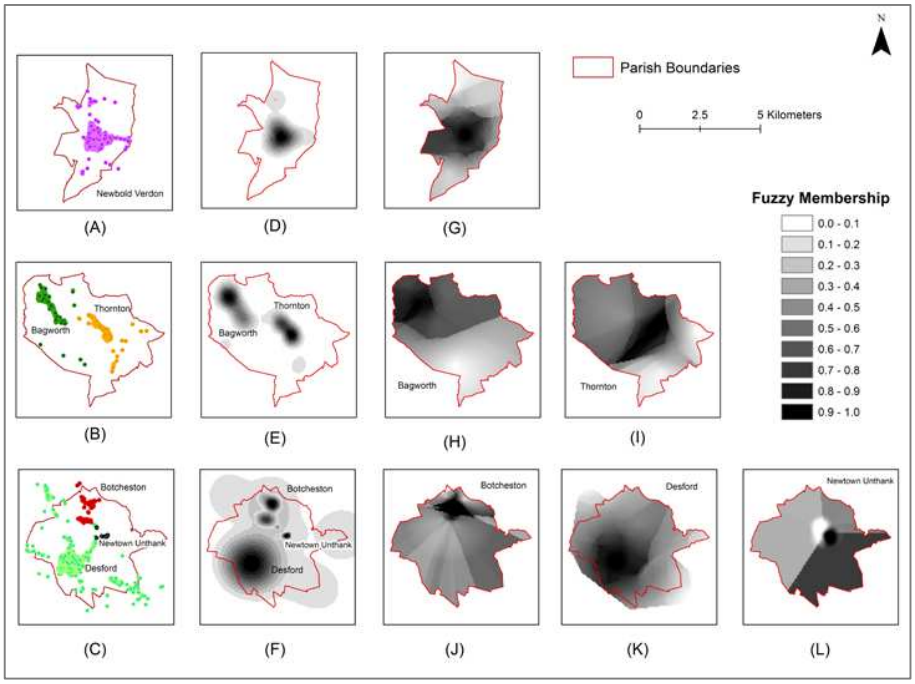



Figure 4: Maps of the source point measurement data (A-C), their normalised density (D-F), and maps of the normalisation of the interpolated surface using ordinary kriging (G-L). The darker the colour, the higher the applicability value at that point to be a member of the settlements.

## 5. Results and Discussion:

A number of observations arose from this work:

- It is clear from Figure 1 that although some settlements have precisely defined parish boundaries (e.g. Desford), the actual extent for the addresses is different; also the same parish may relate to more than one locality (e.g. Botcheston and Newtown Unthank fall in one parish).
- Figure 2 shows that KDE works better with compact zones (areas with high intensity of addresses) than sparsely populated area.
- However, the representation of the fuzzy footprints in Figure 4 is most satisfactory.

Both approaches are normalised to represent the villages as fuzzy objects showing the degree to which a house belongs to a particular village.

These results show the possibility of mapping the geographical extent of villages based on the distribution of addresses in Postal Address records using fuzzy set theory and distance metrics to derive "postal sheds" that define village areas.

## Acknowledgements

The authors would like to acknowledge the MSc work of Tom Gray, which elaborates the main idea of the research. Thanks to the Ordnance Survey for their support and data, as well as special thank to Anne Patrick for facilitating the process. All figures using OS data are © Crown Copyright/database right 2014, an Ordnance Survey/EDINA supplied service. We would also like to take this opportunity to thank King Abdulaziz University and the royal embassy of Saudi Arabia, Saudi Arabian Cultural Bureau in London for their encouragement and financial support.

## References

CHAUDHRY, O. and MACKANESS, W., (2008). Automatic identification of urban settlement boundaries for multiple representation databases. *Computers, Environment and Urban Systems,* **32**, pp.95-109.

CRAWFORD, T.W., (2002). Spatial modelling of village functional territories to support population-environment linkages. In: S.J. WALSH and K.A. CREWS-MEYER, eds, Linking People, Place, and Policy: A GIScience Approach. Boston: Kluwer Academic Publishers, pp.91-111.

Department of COMMUNITIES AND LOCAL GOVERNMENT, (2010). The Local Government Boundary Commission for England. Guidance on community governance reviews. 9781409824213. London: Crown Copyright.

DOWNS, J.A., (2010). Time-Geographic Density Estimation for Moving Point. In: S. FABRIKANT, T. REICHENBACHER, M. KREVELD and C. SCHLIEDER, eds, Geographic Information Science Lecture Notes in Computer Science Volume 6292. Springer Berlin Heidelberg, pp. 16-26.

GOODCHILD, M.F., MONTELLO, D.R., FOHL, P. and GOTTSEGEN, J., (1998). Fuzzy Spatial Queries in Digital Spatial Data Libraries, Fuzzy Systems Proceedings,1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference, 4-9 May 1998 (1998).

GRAY, T., (2008). Modelling a sttlement's spatial extent from placename records using kernel density, convex hull and viewshed analysis. MSc edn. University of Leicester.

HALL, M., SMART, P.D. and JONES, C.B., (2011). Interpreting spatial language in image captions. Cogn Process, 12, pp. 67-94.

HOLLENSTEIN, L. and PURVES, R.S., (2010). Exploring place through user-generated content: using Flickr tags to describe city cores. Journal of Spatial Information Science, 1, pp. 21-48.

KAUFMAN, L. and ROUSSEEUW, P.J., (1990). Finding groups in data: an introduction to cluster analysis. New York: John Wiley and Sons.

LAM, N.S., (1983). Spatial Interpolation Methods: A Review. The American Cartographer, 10(2), pp. 129-149.

LÜSCHER, P. and WEIBEL, R., (2013). Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. *Computers, Environment and Urban Systems,* **37**, pp. 18-34.

MONTELLO, D.R., GOODCHILD, M.F., GOTTSEGEN, J. and FOHL, P., (2003). Where's Downtown? Behavioral Methods for Determinig Referents of Vague Spatial Queries. Spatial Cognition & Computation, 3(2 & 3), pp.185-204.

SILVERMAN, B.W., (1986). Density Estimation for Statistics and Data Analysis. New York, USA: Chapman and Hall.

WILLIAMSON, T., (1996). Vagueness. New York: Routledge.

ZADEH, L.A., (1965). Fuzzy Sets. Information and Control, 8, pp.338-353.

# Where's near? Using web tri-grams to explore spatial relations

Curdin Derungs[1,2] and Ross S. Purves[1]

[1]Department of Geography
University of Zurich, Switzerland
Email: curdin.derungs¦ross.purves@geo.uzh.ch

[2]University Research Priority Program
Language and Space

## 1. Introduction

Vague spatial relations, such as near, are central linguistic building blocks in our everyday discourse (Yao and Thill 2007). As such, it has long been recognised that there is a real need to, firstly, explore and where possible, quantify, how such spatial relations are used, and secondly, develop methods which allow such spatial relationships to be effectively used in information systems.

Approaches to exploring and quantifying the use of spatial relationships have taken, broadly speaking, two distinctive sets of approaches. A number of researchers have explored the use of particular spatial relations through carefully designed empirical experiments with human subjects (e.g. Fisher and Orf 1991; Worboys 2001). These experiments have clearly demonstrated that near is not symmetric, and that context (for example in terms of relative position to other objects, absolute size, or perceived importance) is central in defining whether A is near B (Hernandéz et al. 1995). However, such approaches, though they allow us to hypothesise about potential general properties of vague spatial relations are not scalable to very large numbers of objects.

A similar challenge with respect to the identification of regions associated with vernacular names (Montello et al. 2003) motivated the development of methods to mine such regions automatically from the web (Jones et al. 2008). As methods derived from computer linguistics have become more popular in GIScience, researchers have started to take similar approaches to explore spatial relations in actively and passively crowd-sourced corpora (e.g. Twaroch et al. 2009, Vasardani et al. 2012, Skoumas et al. 2013). These approaches are typically based on a two stage process, firstly identifying and uniquely resolving toponyms and secondly, exploring relations between toponyms, typically through geometry. However, to date most of these studies have still either looked at relatively small regions (e.g. Twaroch et al. explored Cardiff in Wales, Skoumas et al. focused on London), or have used corpora with very specific content (e.g. Vasardani et al. used crowed sourced information from an online game and Skoumas et al. used a crawled corpus consisting of travel blogs).

n-grams are contiguous sequences of n-tokens (e.g. Curdin Derungs or seminal paper are both bigrams) mined from some documents and typically delivered with either frequency or probability values. Both Google and Microsoft have made n-gram collections openly available to researchers, and they are well suited to approaches where a very large corpus may be advantageous, but it is not possible or necessary to process the entire corpus locally (Wang et al. 2010).

In this paper, we aim to extend previous work by, firstly, directly using n-grams to perform our analysis on the spatial relation near used in combination with populated places, secondly, exploring a large region without any restriction on the nature of content (other than that it is found in a web corpus) and, thirdly, making some initial analysis of the influence of context on our results.

## 2. Methodology

Our aim was to derive near relationships for a large number of populated places in a region where we also had access to other data, such as population. We therefore chose to extract all toponyms referring to populated places in Great Britain from GeoNames[1] with more than 1000 inhabitants, and where the toponym was a unigram (n=2050).

Since we expect that many toponyms may be either geo/non-geo (e.g. bath vs Bath) or geo/geo ambiguous (e.g. Washington, Tyne and Wear vs. Washington, D.C.) we first retrieved probabilities for all individual toponyms from the Microsoft n-gram API[2]. The Microsoft n-gram API is openly available for research and returns log10 joint probabilities (effectively the probability of a particular set of tokens appearing within the corpus as an n-gram) based on all documents indexed by Bing in en-us market. For each toponym A we then queried the n-gram API with the trigram A near B, where B was every other toponym in our list. Since we do not expect probabilities for given n-grams to be symmetric, we populate an n x n matrix of probability values and calculated Euclidean distances between toponym pairs.

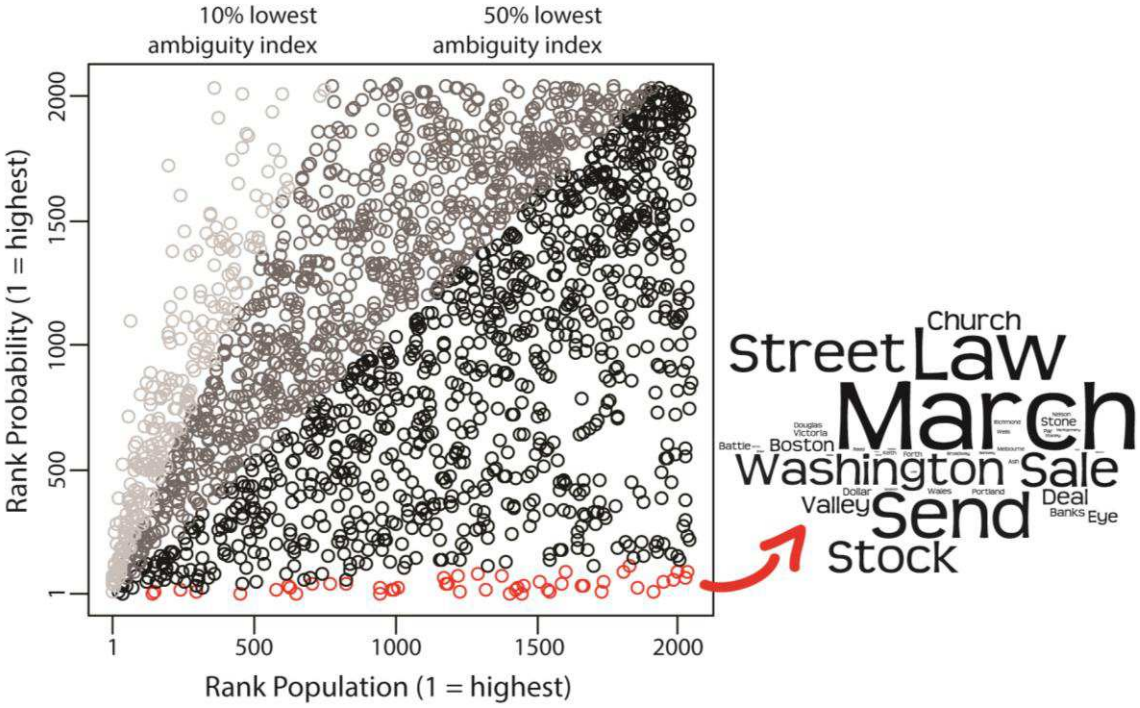



Figure 1. Rank toponym probability against rank population: red toponyms are most ambiguous, and the use of the index to progressively stratify the data is shown in Figure 2

## 3. Results and Interpretation

Figure 1 shows a scatter plot of all selected toponyms, plotted as rank toponym probability against rank population. We would typically expect high population to be associated with more n-grams, and thus a high probability. By exploring the toponyms at the bottom and lower right of the plot (i.e. red dots: high probability but low population) it quickly becomes apparent that these are typically highly ambiguous (e.g. Washington and Boston are both influenced by the American use of the toponym (i.e. geo-geo ambiguous), and Send, March and Over are all geo/non-geo ambiguous.

[1] www.geonames.org
[2] web-ngram.research.microsoft.com/info/api.html

By excluding from our analysis such toponyms, we can explore the influence of ambiguity on our results (a typical problem when using n-grams, as additional data to resolve ambiguity is not available). Thus, we created a simple ambiguity index[3] and used it in further analyses to stratify our data.

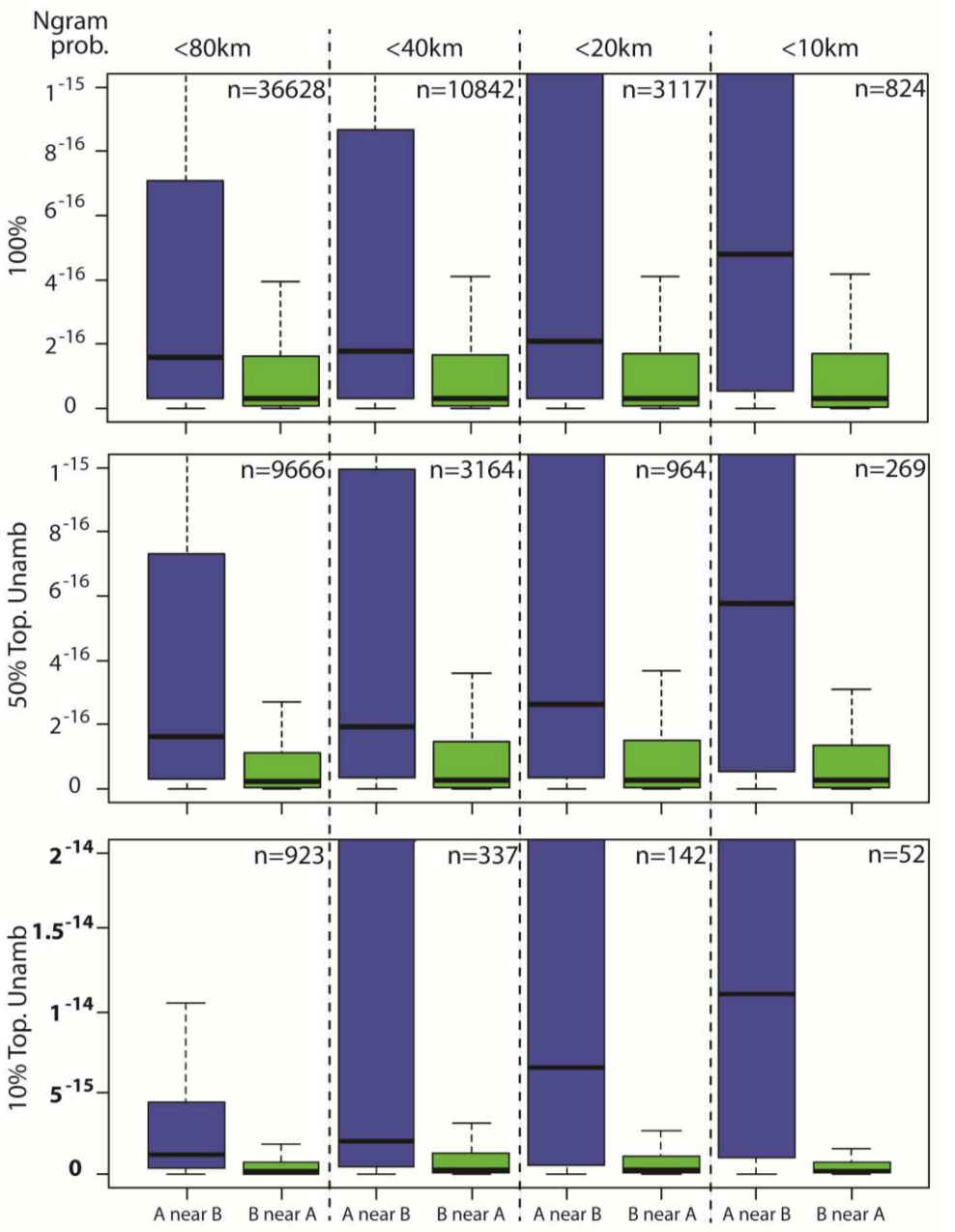



Figure 2. Joint probability for trigram A near B where B has a larger population than A (blue) and where B has a smaller population than A (green) stratified according to Euclidean distance between A and B (columns; using four thresholds of maximum distance) and ambiguity index (rows; filtering toponyms using the ambiguity index)

[3] Effectively the ratio of probability rank over population rank. Importantly, the ambiguity index is only used to demonstrate the effect of ambiguity and not as a disambiguation approach that would allow to analyze the meaning of near.

Figure 2 illustrates bulk results for the whole dataset, stratified according to mean Euclidean distance between toponym pairs (columns), our ambiguity index (rows) and the relative population of the toponyms (boxplot colour). Several observations can be made. Firstly, the joint probabilities of trigrams A near B where the population of B is greater are always greater. This result confirms the importance of context (here population) in defining the asymmetry of nearness relations (c.f. Worboys 2001). Secondly, joint probabilities increase as Euclidean distances decrease demonstrating that, unsurprisingly, nearness is related to distance. Thirdly, ambiguity has a greater influence on joint probabilities than any other factor - considering that the plots representing the 10% top. unambiguous toponyms in Figure 2 has a different scaling of probability values, compared to the two other plots (bolt numbers) -, implying that even simple spatial relationships cannot be explored in such large corpora without accounting for ambiguity. However, although these bulk statistics confirm some simple properties of the use of the term near in language, the variation in joint probabilities demonstrates that simple general rules, such as threshold distances for near, cannot be derived.

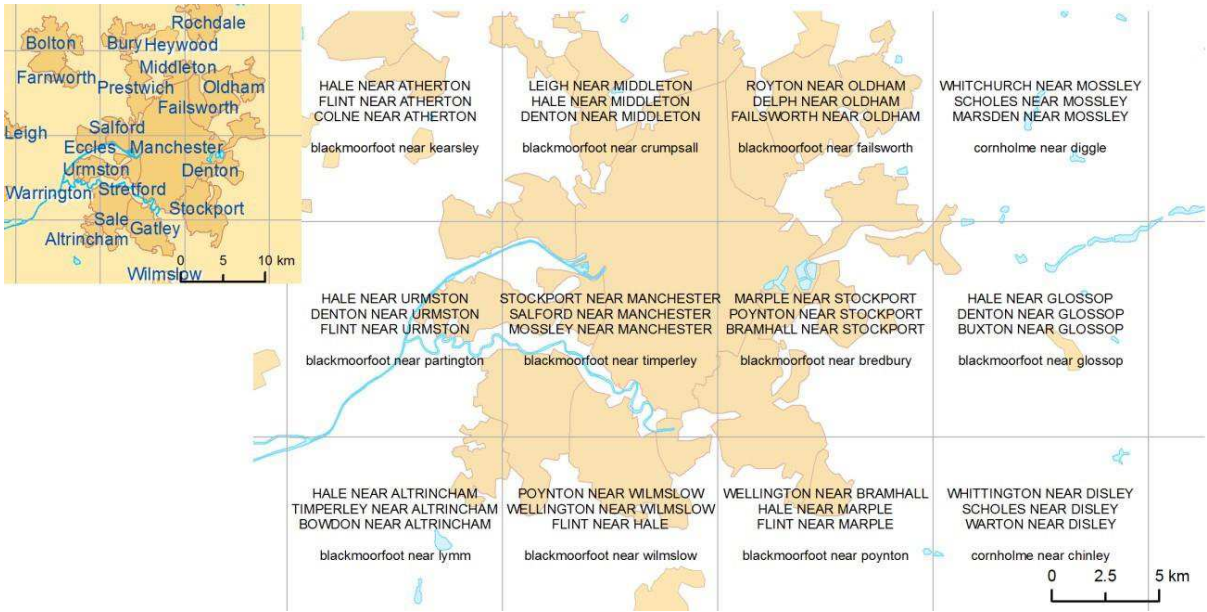



Figure 3. Most probable (upper case) and least probable (lower case) trigrams per grid cell with a 10km resolution (map source: ESRI basemap for Europe).

In Figure 3 we map the spatial pattern of ranked joint probabilities around Manchester. For each 10km grid square we show combinations of A near B, where B is found within the 10km square. A is within 80km of B and B always has a larger population than A. The first three trigrams in each grid square are always the most probable combinations, and the last has the lowest joint probability at this location. A number of observations can be made, which appear to be repeated across the whole of the UK. Firstly, the highest joint probabilities do indeed often seem to be associated with the most prominent B in a grid square (e.g. Manchester, Glossop, Stockport etc.). Secondly, the most probable trigrams appear plausible, and there is evidence of a hierarchy (e.g. Stockport near Manchester; Poynton near Manchester). Thirdly, the least probable trigram regularly repeats toponym A (Blackmoorfoot) suggesting that very low joint probabilities are dominated by the low probability of an individual term. Nonetheless, we suggest that this map illustrates that we can move towards mining meaningful trigrams representing nearness within a particular region, which could be used as a tool for resolving ambiguity.

## 3. Conclusions

Our results clearly demonstrate that n-grams have considerable potential, but that the advantage offered by being able to query simple tri-grams (instead of having access to whole documents) also brings challenges in understanding potential ambiguity and local context. We were able to illustrate cases where ambiguity is probable, but this approach forces us to discard up to 90% of our initial data set. Furthermore, this ambiguity appeared to have more influence on the joint probabilities we generated than any other form of stratification, again demonstrating that ignoring it (as is currently the case in many studies with large corpora and very simple text mining approaches) is not an option. However, stratifying the data allows us to explore, and identify, relationships which are sensible, such as the influence of context on nearness.

One strength of our approach is its generic character, such that it can be applied to different spatial scales and extended to a variety of spatial relations (e.g. directional or topological). However, we suggest that automatically deriving definitions of spatial relations from these data requires further research, but n-grams provide a rich resource to explore the extent how geography matters in spatial relationships.

## Acknowledgements

We would like to thank the University Research Priority Program Language and Space of the University of Zurich for supporting this work and Microsoft for giving access to the ngram data.

## References

Fisher PF and Orf TM, 1991, An investigation of the meaning of near and close on a university campus. Computers, Environment and Urban Systems, 15(1):23–35.

Hernandez D, Clementini E and Di Felice P, 1995, Qualitative distances. Springer.

Montello D, Goodchild M, Gottsegen J and Fohl P, 2003, Where's Downtown? Behavioral Methods for Determining Referents of Vague Spatial Queries. Spatial Cognition & Computation, 3(2):185-204.

Skoumas G, Pfoser D, Anastasios TK, 2013, On quantifying qualitative geospatial data: a Probabilistic approach, In: Proceedings of the 2nd GEOGROWD Workshop, 71-78.

Twaroch F, Purves RS and Jones C, 2009, Stability of Qualitative Spatial Relations between Vernacular Regions Mined from Web Data, In: Proceedings of Workshop on Geographic Information on the Internet, Toulouse, France.

Vasardani M, Winter S, Richter KF, Stirling L and Richter D, 2012, Spatial Interpretations of Preposition "at", 2012 ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information (GEOCROWD), ACM Press, Redondo Beach, CA.

Wang K, Thrasher C, Viegas E, Li X and Hsu BP, 2010, An overview of Microsoft Web N-gram corpus and applications. In: Proceedings of the NAACL HLT 2010 Demonstration Session, 45–48.

Worboys MF, 2001, Nearness relations in environmental space. International Journal of Geographic Information Science, 15(7):633-651.

Yao X, and Thill JC, 2007, Neurofuzzy modeling of context--contingent proximity relations. Geographical analysis, 39(2):169–194.

# Eye tracking with geographic coordinates: methodology to evaluate interactive cartographic products

Kristien Ooms[1]; Sara Fabrikant[2] Arzu Coltekin[2]; Philippe De Maeyer[1]

[1]Department of Geography; Ghent University, Belgium
{kristien.ooms; philippe.demaeyer}@ugent.be
[2]Departement of Geography; University of Zurich, Switzerland
{sara.fabrikant; arzu.coltekin}@geo.uzh.ch

## 1. Introduction

Empirical research (e.g., User Centred Design or UCD) has repeatedly shown that involving users early on in a product's iterative design has led to major improvements in its usability, but UCD's effective implementation might be still cumbersome (Nivala et al. 2007). One of the main challenges is the balance between ecological validity and experimental control. Data collected in controlled lab studies might be more straightforward to process, but results might not reflect real world use situations. The latter may suffer from hard to control (potentially confounding) variables, unpredictable test conditions, and thus less consistent and comparable study outcomes across participant groups (e.g. Nielsen 1993). This delicate balance can also be found in cartographic user research, for example, when employing the eye tracking methodology to register and analyse users' overt visual behaviour (e.g. Coltekin et al. 2009; Fabrikant et al. 2008; Ooms et al. 2012). In general, current state-of-the-art eye tracking systems have limited automated solutions to deal with the analysis of interactive stimuli. Moreover, users' gaze locations (or Points of Regard, POR), are typically recorded in screen coordinates (e.g., pixel locations in a display) and not in geographic coordinates, which introduces a spatial data analysis challenge when evaluating interactive cartographic products. Nevertheless, the viewed geographic locations might be particularly relevant for a specific spatial decision making task.

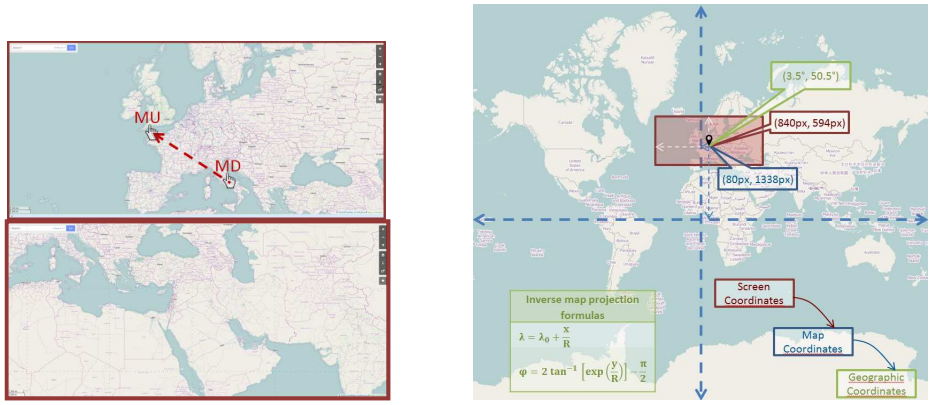
Interactive maps in user studies are often approximated by pre-computed animations or by automatically loading a number of subsequent static images (e.g. Fabrikant et al. 2008; Ooms et al. 2012). In doing so, the experimenter introduces a high level of experimental control to facilitate empirical data analysis with dynamic displays. To increase ecological validity, however, participants should be able to execute a task on interactive maps as they would normally do, that is, without restricting their inference making behaviour or the interactivity levels of the tested map display. In the next section, we propose a user-centred evaluation framework based on the eye tracking methodology coupled with user logging to specifically evaluate a wide range of interactive cartographic products.

## 2. Georeferencing eye movements on interactive maps?

To evaluate interactive cartographic products, it is essential that human-map interactions are tracked. In UCD, user-system interaction logging (e.g., mouse movements, key-stroke analyses, etc.) is often utilized to gather quantitative data from users who execute a task with a product (Haklay & Nivala 2010; Nielsen 1993; Slocum et al. 2001; van Elzakker & Griffin 2013), and this has also been coupled with eye tracking on interactive maps (Coltekin et al. 2009). Depending on the employed eye tracker, low-level user logging might not be readily available, thus additional logging software is typically needed to record detailed user-system interactions (Coltekin et al. 2009). From the range of available loggers, we chose the open source PyHook library to develop custom scripts that *hook into* the computer's operating system, and record detailed user interactions: mouse movements, mouse key presses and releas-

es, and keyboard actions. This desktop-based library works independently from the eye tracker and the digital map application. Consequently, nearly any map product can be evaluated, whether it is a third-party, online map mashup (independent of its API), or any type of (offline) desktop mapping applications.

Logging user actions together with eye movements makes it possible to not only determine when and how user interactions occur, but it also allows to capture where exactly on the map a participant was looking at a certain moment in time. For geographic analyses, this collected data should ideally be in map or geographic coordinates. We detail below how this can be achieved for the panning operation.




(a) (b)

Figure 1. (a) Illustration of the pan operation and (b) different locational reference systems (screen, map, and geographic coordinates)

Panning operations can be conceptualized as a fixed window frame that a user moves over a map without changing the map's viewing scale (red rectangle in Figure 1). This operation is defined by a *mouse key down* (MD) event and consecutive *mouse key up* (MU) event (Figure 1a). By default, the users' POR is registered in screen coordinates, thus relative to the upper left corner of the red rectangles in Figure 1a. As the map scale does not change during the pan event, it is possible to transform the screen coordinates of the POR into the respective map coordinates, relative to the map's centre (blue coordinates in Figure 1b). If the geographic reference system and map projection parameters are known, map coordinates can be transformed into geographic coordinates. Regarding popular online mapping platforms, such as OpenStreetMap, it is the WGS84 locational reference system and a spherical Mercator projection. Using the inverse map projection formula one can re-calculate recorded map pixel locations in the current viewing window to geographic coordinates (green coordinates in Figure 1b).

## 3. Case study

We employed the OpenStreetMap (OSM) web mapping platform for our proof-of-concept study, and recorded users' POR during three test sessions, each with one of the three most used eye tracking systems: SMI RED250, Tobii T120, and SR Research's EyeLink1000. We followed the identical test protocol. After calibrating participants with the eye tracker, they were asked to press a button to synchronise the internal clock of the eye tracker with the PyHook logger. The screen recording mode was then started to record the entire test session. The same OSM URL was loaded into the Web browser window to make sure that all participants started viewing the map at the same scale and in the same geographic region (i.e. top

left image in Figure 2). Participants were then asked to pan to other world regions as illustrated in Figure 2.

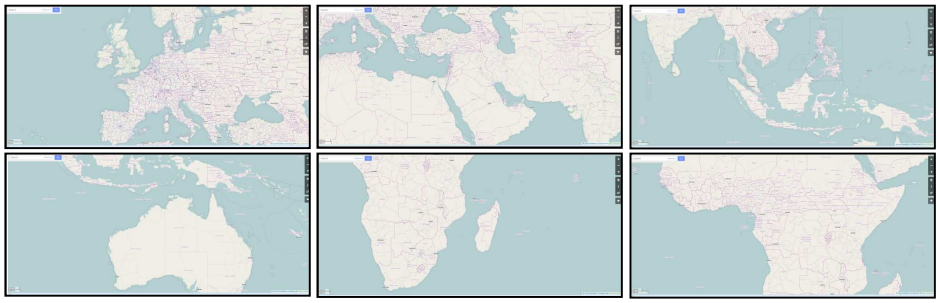


Figure 2. Test stimuli and task: pan operation to different world regions

Participants' raw eye movement records (in screen coordinates) were aggregated into fixations using the analysis software associated with each of the used eye trackers. The recorded screen coordinates were transformed into OSM-map coordinates (related to scale level 5), and then to spherical geographical coordinates, as detailed above. Resulting fixation locations were then imported into ArcGIS and visualized over a static world map with OSM's map projection (Figure 3). Visually comparing the target regions (in Figure 2) with the fixated locations in Figure 3, it appears that the fixations recorded with all three systems are indeed located in the expected regions.

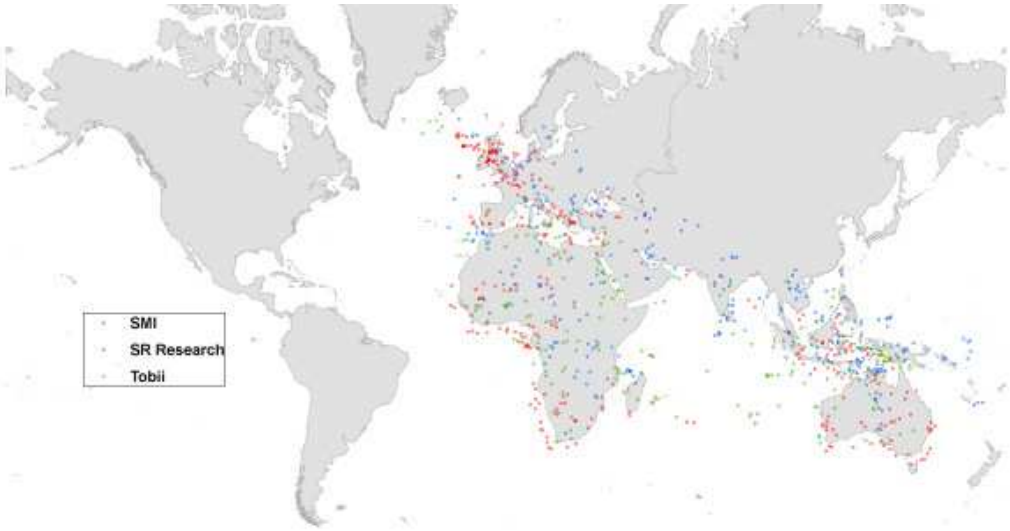



Figure 3. Participants' fixations recorded with different eye trackers

However, when synchronizing recorded time stamps from the eye tracker with those from the logging tool, we discovered small deviations (of maximum 10 ms) between the two. This is not uncommon, and can have various reasons (e.g., different internal clock settings, influence of computer processor speed, etc.). Nevertheless, it still acceptable for our purposes, taking into account the minimal eye tracker sampling rate: a gaze location recorded every 8.33ms (i.e., SMI and Tobii).

## 4. Conclusion

We propose a user-centred evaluation framework for interactive cartographic products using eye tracking coupled with automated user logging that transforms recorded eye movement data from interactive map stimuli (expressed in screen coordinates) to map coordinates and/or

spherical geographic coordinates. The resulting eye records coupled with the interaction data (e.g., gaze locations before/after an interaction) can be analysed using different coordinate systems in a GIS (e.g., where on the screen, on the map or in the world was the user looking?). Georeferenced gaze data allows further spatial data processing, using straightforward spatial analysis techniques readily available in off-the-shelf GIS (e.g., buffering, cluster detection, etc.). We believe that our proposed approach will greatly facilitate the empirical study of interactive map use and human decision making with digital maps.

## References

Coltekin, A., Heil, B., Garlandini, S., & Fabrikant, S. I. (2009). Evaluating the effectiveness of interactive map interface designs: a case study integrating usability metrics with eye-movement analysis. *Cartography and* Geographic *Information Science, 36*(1), 5-17.

Fabrikant, S. I., Rebich-Hespanha, S., Andrienko, N., Andrienko, G., & Montello, D. R. (2008). Novel method to measure inference affordance in static small-multiple map displays representing dynamic processes. *The* Cartographic *Journal, 45*(3), 201-215.

Haklay, M., & Nivala, A. M. (2010). User-Centred Design. *Interacting with geospatial technologies*, 89-106.

Nielsen, J. (1993). *Usability Engineering*. San Francisco: Morgan Kaufmann.

Nivala, A.-M., Sarjakoski, L. T., & Sarjakoski, T. (2007). Usability methods' familiarity among map application developers. International *Journal of Human-Computer Studies, 65*, 784-795.

Ooms, K., De Maeyer, P., Fack, V., Van Assche, E., & Witlox, F. (2012). Interpreting maps through the eye of expert and novice users. *International Journal of Geographical Information Science, 26*(10), 1773-1788.

Slocum, T. A., Blok, C., Jiang, B., Koussoulakou, A., Montello, D. R., Fuhrman, S., & Hedley, N. R. (2001). Cognitive and usability issues in geovisualisation. *Cartography and Geographic Information Science, 28*(1), 61-75.

van Elzakker, C. P. J. M., & Griffin, A. L. (2013). Focus on Geoinformation Users: Cognitive and Use/User Issues in Contemporary Cartography. *GIM International, 27*(8), 20-23.

# Inferring population structure from surname geography: The role of GIScience in interpreting genetic information

Jens Kandt[1 2], James A. Cheshire[4], Paul A. Longley[1 3]

[1] Department of Geography, University College London, Gower Street, London WC1E 6BT, UK
[2] Email: j.kandt.12@ucl.ac.uk
[3] Email: p.longley@ucl.ac.uk
[4] UCL Centre for Advanced Spatial Analysis, University College London, Gower Street, London WC1E 6BT, UK
Email: james.cheshire@ucl.ac.uk

## 1. Introduction

Thanks to the increasing availability of DNA sequenced data, the genetic structure of populations can now be studied at a fine resolution. This possibility is not only interesting for historical and genetic research in general, but it is also relevant for the design of biomedical studies investigating associations between genomes and diseases (Winney et al 2012). Yet genetic samples are special because the precise population they represent is uncertain, not to say unknown. Although this problem may not be significant in world-wide population studies (Cavalli-Sforza 1994), the question of representativeness at a finer scale determines the possibility of valid inferencing for the two above-mentioned research purposes. In this paper, we present geographical research methods to relate a British genetic sample to the general population in Britain and argue that GIScience is crucial to enable the use of emerging high resolution genetic information.

## 2. Data and methods

We used pre-analysed DNA sequences of 2,019 British volunteers who participated in a Wellcome Trust-funded study. In order to best represent local sampling pools, volunteers were included if their grandparents' birthplaces were in rural Britain and no more than 80 kilometres apart from each other (see Winney et al 2012 for details). The mean birth year of grandparents was 1885.

We took surnames of the 1881 Census of England Wales and Scotland micro-dataset and the 2007 GB Enhanced Electoral Roll, both of which can be regarded as complete population registers. We calculated lcoal concentrations of surnames in 1881 parishes and 2007 wards. Local surname concentrations naturally correlate with the geographical distribution of genetic variants (Darlu et al 2012; Degioanni et al 2003; Jobling 2001). We used a so-called coancestry matrix to measure genetic similarities between each pair of volunteers. For surnames, we calculated an isonymy matrix (Lasker 1977), which measures the similarity of surname compositions of each pair of local areas (parishes or wards). We excluded urban areas and foreign surnames. The two matrices can be transformed into dissimilarity or distance matrices, which are usable in cluster algorithms..

## 3. Global consonance: inferring fine population structure

In order to measure the global association between surnames and gene frequencies, the transformed isonymy matrix was processed in Ward's hierarchical clustering algorithm (Everitt 1974) to produce between 2 to 80 clusters of areas with distinct surname compositions. For the genetic data, we used cluster solutions with 2 to 53 clusters developed by Lawson and colleagues (2012). First, we determined for each volunteer the genetic cluster membership for each of the $l$ cluster solutions ($l \in \{2..53\}$). Then, we appended to the volunteer information the membership of each $k$ surname cluster ($k \in \{2..80\}$) based on the

location of the grandparents' birth places. We subsequently measured the level of agreement between various *l* and *k* using the Adjusted Rand (AR) similarity index (Albatineh et al 2006).

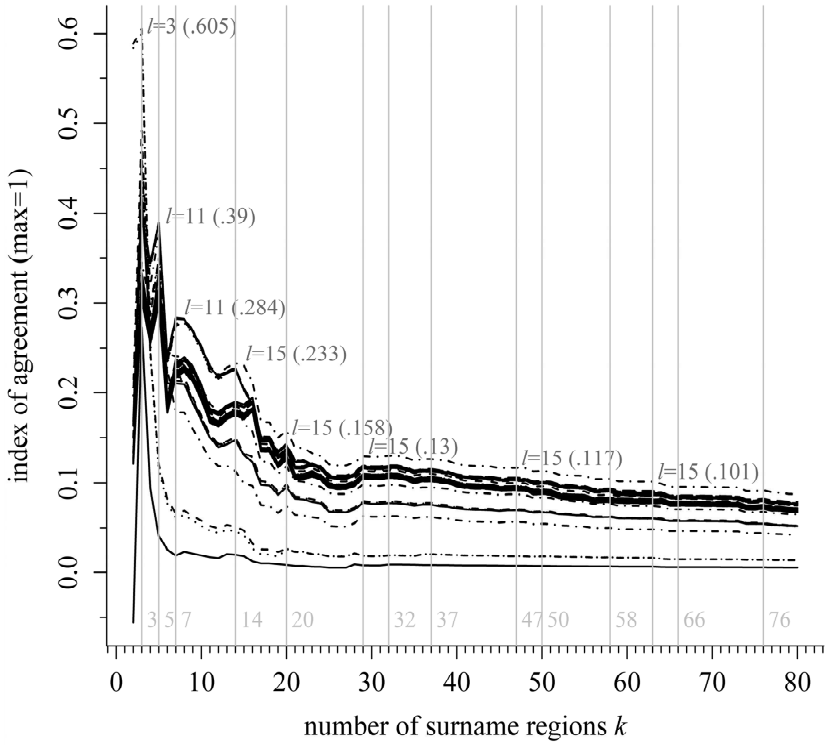



Figure 1. Adjusted Rand Index values of agreement between *l* genetic clusters and *k* surname regions. Each line represents a genetic cluster solution with $l \in \{2..53\}$.

The index shows a global peak at $l = 3$ genetic clusters and $k = 3$ surname clusters (Fig. 1). The next best local peak is at $l = 11$ genetic clusters and $k = 5$ surname regions, followed by other local peaks at $l = 11$ and $k = 7$, $l = 15$ and $k = 14$, $l = 15$ and $k = 20$ and so forth.

Mapping the most-agreeing cluster solutions reveals strong spatial concentration of the clusters (Fig. 2). It should be noted that geographic information was not processed in the clustering. The surname cluster solution with $k = 3$ divides Great Britain into three regions: Wales, England and Scotland. In the genetic cluster solution with $l = 3$, Wales, England-Scotland and Orkney emerge as regions. With an AR of 0.61, this combination indicates high certainty about real population structure.

Figure 3 provides an example at a much finer level, comparing 20 surname regions and 15 genetic clusters. The maps show similar regionalisations in West and East Scotland, North England, Wales and at its borders, South East England, as well as Cornwall and Devon in the South West. But we can also identify parts of Britain, where genes and surnames differ: in Orkney, for example, we find more genetic heterogeneity than surnames suggest, and vice versa, in South East England, we find more surname diversity than genetic diversity.

(a) $k = 3$ surname regions (b) $l = 3$ genotypes

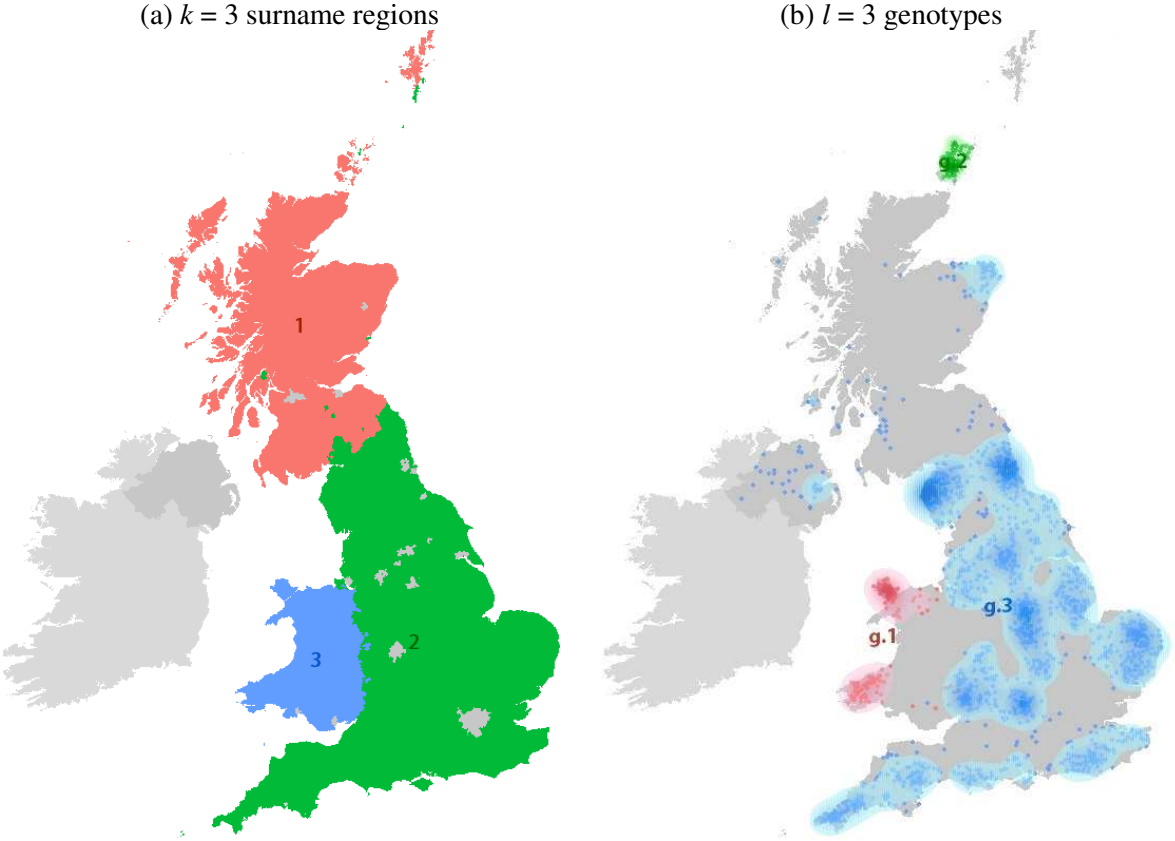



Figure 2. Three surname regions (a) and three genetic clusters (b) with AR = .605.

(a) $k = 20$ surname regions (b) $l = 15$ genotypes

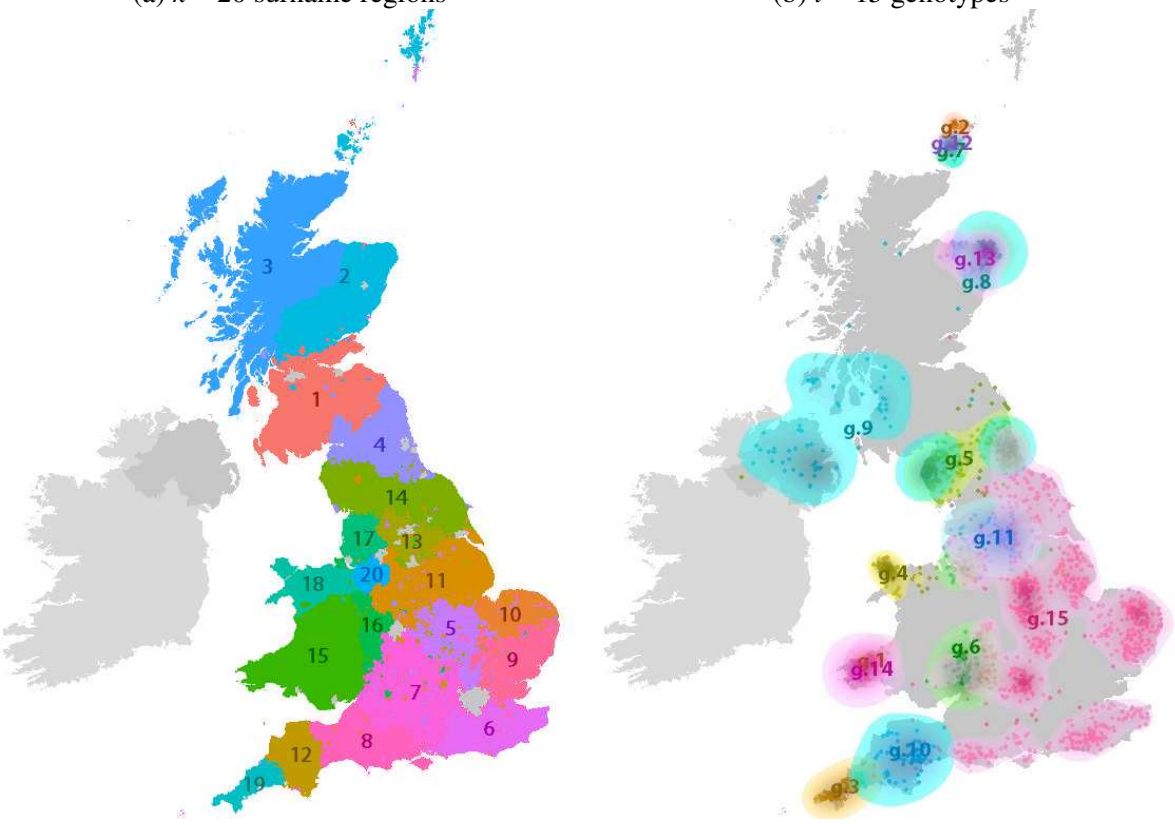



Figure 3. 20 surname regions (a) and 15 genetic clusters (b) with AR = .158.

## 4. Local dissonance: informing sample design

Multi-Dimensional Scaling (MDS) can be used to summarise genetic characteristics of individuals and surname compositions of areas in up to three dimensions (Herrera Paz 2014). We mapped and interpolated the MDS dimension scores geographically using Inverse Distance Weighting (IDW). The MDS surfaces exhibit strong geographical patterning of both genetic similarity and surname compositions (Fig. 4).

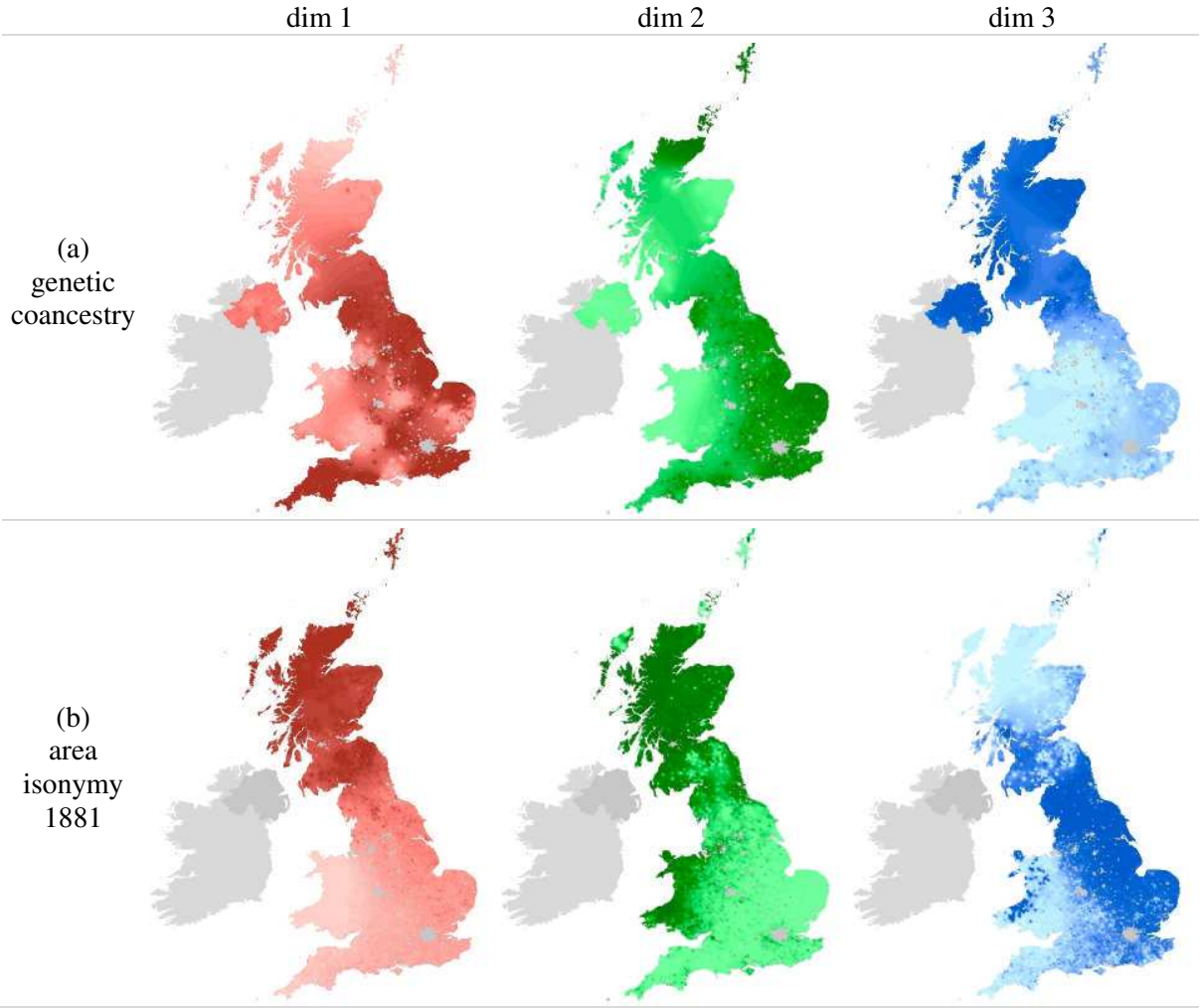



Figure 4. The three dimensions of multidimensional scaling of genetic coancestry (a) and area isonymy (b) based on local 1881 surname compositions. Darker colours indicate similarity.

Populations of Wales and Scotland emerge as different from English populations in terms of both surnames and genes, with increasing differences as geographic distance increases. The geographical patterns of both surnames and genes are similar, albeit that the MDS dimensions appear in different orders and can be reversed.

In order to measure local consonance or dissonance between surnames and genes, the surfaces of each dataset were correlated and the residuals were plotted in another IDW (Fig. 5). The higher the residuals are, the higher the dissonance between coancestry and isonymy. In regions with high residuals (darker coloured) on one dimension, we would be less certain about true differences in population structure. This information is useful for deciding on control and contrast samples, for example in epidemiological studies. The patterns remain broadly similar in 1881 and 2007.

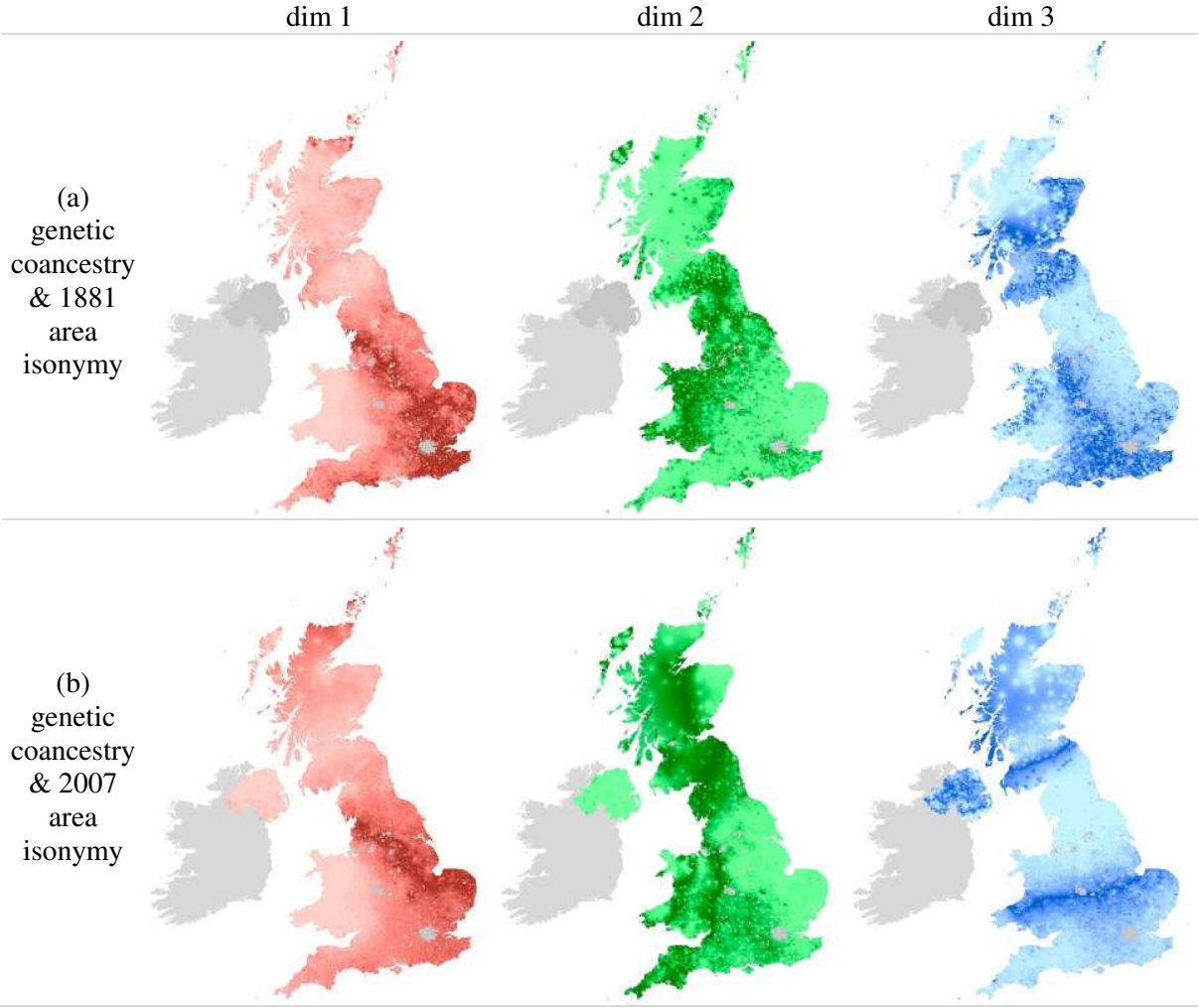



Figure 5. Residuals of correlated MDS dimensions between genetic coancestry and area isonymy in 1881 (a) and 2007 (b). The darker the colours, the higher are the residuals indicating local dissonance between genes and surnames.

.

## 5. Conclusions: the role of GIScience

GIScience is a crucial instrument to leverage the potential of genetic data for various research domains. The already geographically constrained design of the genetic sample permits the application of GIScience as a spatial heuristic to relate genetic information to the whole population and support the inference of fine population structure at different levels of uncertainty. The results suggest that applied GIScience can sharpen aetiological disease research, such as genome-wide association studies, and improve chances of valid inferencing through more robust biomedical sampling geographies.

## Acknowledgements

This study is part of the larger project 'The people of the British Isles' funded by the Wellcome Trust, UK. We would like to thank Bruce Winney, Garrett Hellenthal and Walter Bodmer for provision of co-ancestry matrix and genetic cluster assignments. We are indebted to Humphrey Southall and Paula Aucott, for making available supporting information for the 1881 Census from the Vision of Britain project.

## References

Albatineh, A.N., Niewiadomska-Bugaj, M. & Mihalko, D., 2006. On Similarity Indices and Correction for Chance Agreement. *Journal of Classification*, 23(2), pp.301–313.

Cavalli-Sforza, L., Menozzi, P. & Piazza, A., 1994. *The history and geography of human genes* P. Menozzi & A. Piazza, eds., Princeton, N.J: Princeton, N.J : Princeton University Press.

Darlu, P., Bloothooft, G. & Boattini, A., 2012. The family name as socio-cultural feature and genetic metaphor: From concepts to methods. *Human Biology*, 84(2), pp.169–214.

Degioanni, A., Darlu, P. & Raffoux, C., 2003. Analysis of the French National Registry of unrelated bone marrow donors, using surnames as a tool for improving geographical localisation of HLA haplotypes. *European journal of human genetics : EJHG*, 11(10), pp.794–801.

Everitt, B. 1974. *Cluster Analysis*. London: Heinemann Educational for the Social Science Research Council.

Herrera Paz, E. F., Scapoli, C., Mamolini, E., Sandri, M., Carrieri, A., Rodriguez-Larralde, A., & Barrai, I. (2014). Surnames in Honduras: A study of the population of Honduras through isonymy. *Annals of Human Genetics* 78(3), 165–77

Jobling, M.A., 2001. In the name of the father: surnames and genetics. *Trends in genetics : TIG*, 17(6), pp.353–357.

Lasker, G., 1977. A coefficient of relationship by isonymy: a method for estimating the genetic relationship between populations. *Human Biology*, 49(3), pp.489–493.

Lawson, D.J. et al., 2012. Inference of population structure using dense haplotype data. *PLoS genetics*, 8(1), pp.1-16.

Winney, B. et al., 2012. People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *European journal of human genetics : EJHG*, 20(2), pp.203–210.

# Dasymetric Refinement for Improved Temporal Small Area Analysis

S. Leyk[1], B. P. Buttenfield[1], M. Ruther[2]

[1]University of Colorado Boulder, Boulder, Colorado, USA
Email: { stefan.leyk, babs }@colorado.edu

[2]University of Louisville, Louisville, Kentucky, USA
Email: matthew.ruther@louisville.edu

## 1. Introduction

Temporal analysis of small area demographic estimates improves understanding about patterns of population change and can generate new knowledge about social and demographic processes over time. However, comparability among demographic data derived from different censuses is compromised by changing enumeration boundaries over time. In preparing data for temporal analysis, zone incompatibilities are commonly solved using areal interpolation methods such as Areal Weighting (Markoff and Shapiro 1973; Goodchild and Lam 1980) or Target Density Weighting (Schroeder 2007). Such techniques are limited by the underlying assumption of homogeneous density within each source and target zone, which is often inaccurate in describing the distribution of human populations. In cases where census boundaries change in response to population change, this assumption can lead to large estimation errors (Gregory 2002).

Areal interpolation methods may integrate ancillary data in a form of dasymetric refinement to improve the accuracy of estimation for a single point in time (Eicher and Brewer 2001; Mennis and Hultgren 2006; Langford 2007; Lin et al. 2011). Dasymetric refinements are commonly based on remotely sensed imagery or classified land cover data (Mennis 2003; Holt et al. 2004; Kim and Yao 2011) or street network data (Reibel and Bufalino 2005). Areal interpolation methods that incorporate dasymetric refinement have been shown to lead to more accurately interpolated small area estimates, relative to methods that do not use ancillary data (Mrozinski and Cromley 1999; Gregory 2002). Dasymetric methods also tend to out-perform regression-based statistical models (Fisher and Langford 1995; Lin et al. 2011).

Recent examples of areal interpolation for temporal analysis use ancillary information to adjust weights for assigning population counts to different land cover classes (Fisher and Langford 1995; Cockings et al. 1997; Holt et al. 2004; Reibel and Bufalino 2005; Reibel and Agrawal 2007; Schroeder and van Riper 2013). Rarely is ancillary information used however for the direct adjustment of the units of analysis. Notably absent from this literature is a comparison of different areal interpolation methods that incorporate dasymetric refinement, or an evaluation of the contexts under which spatial refinement may be beneficial. This research explores the improvement of currently adopted methods to resolve zonal incompatibilities over time using dasymetric refinement prior to interpolation.

## 2. Data and Methodology

The analysis entails four urban counties which reflect a variety of population changes over the 2000-2010 study period: rapid growth (Clark County, Nevada containing Las Vegas), modest growth (Hennepin County, Minnesota containing Minneapolis), modest decline (Allegheny County, Pennsylvania surrounding Pittsburgh), and rapid decline (Wayne County, Michigan, containing Detroit).

The unit of analysis is the census tract, a U.S. Census-based small enumeration area commonly used in demographic analysis. Census tract boundaries in 2010 serve as target zones, and population estimates are interpolated for these target zones using source zone Census data from 2000. Binary dasymetric refinement is supported by data from the National Land Cover Database (NLCD 2001 & 2011) (Figure 1). Pixels are classified as developed land (NLCD classes 21-23) or non-developed land (all other classes).


Figure 1. Developed land as classified in NLCD 2001 (left panel) and NLCD 2011 (right panel) used as ancillary variable for dasymetric refinement, shown within Census 2000 and 2010 tract boundaries, respectively.

Four areal interpolation methods compare Areal Weighting (AW); Tobler's Pycnophylactic Method (PM) (Tobler 1979); Target Density Weighting (TDW) (Schroeder 2007); and the Expectation-Maximization (EM) method (Dempster 1977; Flowerdew and Green 1994). For the first three methods, interpolation will be carried out twice: once using no dasymetric refinement and a second time using a binary dasymetric refinement. Refinement will be applied for AW and PM in the source year only and for TDW in both years. Dasymetric refinement is inherent in the EM model.

Interpolation error will be quantified using Census block data from 2000, which approximately nests within both the 2000 source zones and the 2010 target zones. Error metrics will be computed within each county for those Census tracts which exhibit substantial boundary changes between 2000 and 2010: the mean and median error, the root mean square error (RMSE), and the 90th percentile error.

## 3. Initial Results

Table 1 compares errors from unrefined interpolation with the dasymetrically refined interpolation in estimating 2000 population counts within 2010 target zones. The refined TDW method outperforms the other methods by nearly all metrics in each of the study areas. The errors are highest overall in Las Vegas. This is likely the result of Clark County exhibiting extreme boundary changes due to rapid growth during the 2000-2010 decade.

Table 1. Error metrics for population estimates using unrefined and dasymetrically refined (DR) areal interpolation

| | | AW | AW-DR | TDW | TDW-DR | PM | PM-DR | EM |
|---|---|---|---|---|---|---|---|---|
| **Wayne** | Mean | 381 | 257 | 212 | 167 | 352 | 278 | 257 |
| N = 79 | Median | 112 | 102 | 66 | 60 | 98 | 99 | 101 |
| (out of 609) | 90% | 1142 | 662 | 646 | 451 | 1062 | 801 | 874 |
| | RMSE | 746 | 476 | 411 | 301 | 749 | 551 | 475 |
| **Hennepin** | Mean | 307 | 249 | 205 | 189 | 311 | 254 | 226 |
| N = 53 | Median | 110 | 116 | 90 | 104 | 116 | 124 | 129 |
| (out of 299) | 90% | 735 | 720 | 540 | 539 | 784 | 770 | 447 |
| | RMSE | 507 | 364 | 354 | 305 | 511 | 395 | 361 |
| **Allegheny** | Mean | 326 | 249 | 127 | 101 | 327 | 259 | 199 |
| N = 151 | Median | 59 | 36 | 73 | 57 | 58 | 41 | 49 |
| (out of 402) | 90% | 1088 | 974 | 317 | 227 | 1131 | 812 | 627 |
| | RMSE | 775 | 627 | 205 | 186 | 833 | 663 | 453 |
| **Clark** | Mean | 741 | 546 | 457 | 356 | 604 | 490 | 428 |
| N = 241 | Median | 448 | 309 | 313 | 213 | 371 | 249 | 274 |
| (out of 487) | 90% | 1808 | 1240 | 1148 | 880 | 1490 | 1335 | 1010 |
| | RMSE | 1173 | 903 | 673 | 535 | 954 | 776 | 684 |

Dasymetric refinement results in substantial reductions in the mean error and the RMSE for all methods in all study areas; the median error and 90$^{th}$ percentile error are also nearly universally reduced. Figure 2 displays the distribution of interpolation errors from TDW and dasymetrically refined TDW for the target zones in central Clark County. Largest interpolation errors are visible in the newly developing tracts at the perimeter of the county.

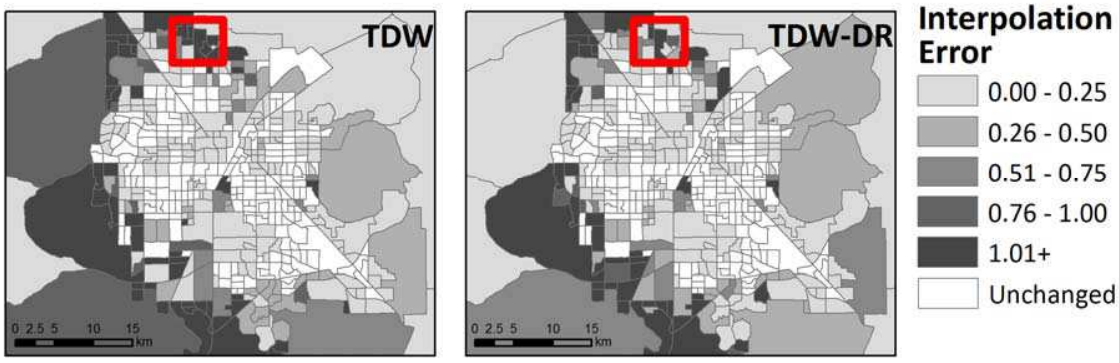



Figure 2. Interpolation errors using TDW and TDW-DR in Clark County, NV. The red box shows the area highlighted in Figure 3.

Figure 3 shows the same geographic footprint in Clark County as in Figure 1, and highlights how the improvements in performance between TDW and the dasymetrically refined TDW appear to be related to differences in land cover.


Figure 3. Interpolation errors for unrefined and refined TDW

## 4. Next Steps

The accuracy of any particular areal interpolation technique is likely to be affected by the underlying population geographies in the source and target zones, as well as by relationships identified between the population surfaces and the ancillary data (Mennis and Hultgren 2006). To identify the sources of areal interpolation error, a regression model of estimation error will be constructed similar to Schroeder's (2007), in which different boundary and tract properties are used as predictive variables.

Also, the time period will be extended to include additional census years and thus provide a richer basis for temporal analysis. Future research will also improve dasymetric refinements using additional ancillary variables to further reduce the interpolation errors.

## Acknowledgements

Funded by the National Science Foundation: "Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata", Project BCS-0961598 awarded to University of Colorado – Boulder.

## References

Cockings S, Fisher PF, and Langford M. 1997. Parameterization and visualization of the errors in areal interpolation. *Geographical Analysis* 29(4):314-328.

Dempster AP, Laird NM, and Rubin DB. 1977. Maximum likelihood for incomplete data via the EM algorithm. *Journal Royal Statistical Society B* 39:1-38.

Eicher CL and Brewer CA. 2001. Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science* 28(2):125-138.

Fisher PF and Langford M. 1995. Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A* 27(2):211-224.

Flowerdew R and Green M. 1992. Statistical methods for inference between incompatible zonal systems. In: Goodchild MF and Gopal S (Eds.). *Accuracy of Spatial Databases.* London: Taylor and Francis.

Goodchild MF and Lam MS. 1980. Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing* 1:297-331.

Gregory IN. 2002. The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems* 26:293-314.

Holt JB, Lo CP, and Hodler TW. 2004. Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science* 31(2):103-121.

Kim H and Yao X. 2010. Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method. *International Journal of Remote Sensing* 31(21):5657-5671.

Langford M. 2007. Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems* 31:19-32.

Lin J, Cromley R, and Zhang C. 2011. Using geographically weighted regression to solve the areal interpolation problem. *Annals of GIS* 17(1):1-14.

Markoff J and Shapiro G. 1973. The linkage of data describing overlapping geographical units. *Historical Methods Newsletter* 7:34–46.

Mennis J and Hultgren T. 2006. Intelligence dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science* 33(3):179-194.

Mennis J. 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer* 55(1):31-42.

Mrozinski RD and Cromley RG. 1999. Singly – and doubly – constrained methods of areal interpolation for vector-based GIS. *Transactions in GIS* 3(3):285-301.

Reibel M and Agrawal A. 2007. Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review* 26:619-633.

Reibel M and Bufalino ME. 2005. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A* 37(1):127-139.

Schroeder JP and Van Riper DC. 2013. Because Muncie's densities are not Manhattan's: using geographical weighting in the expectation-maximization algorithm for areal interpolation. *Geographical Analysis* 45:216-237.

Schroeder JP. 2007. Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis* 39:311-335.

Tobler WR. 1979. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association* 74(367):519-530.

# Normalized Mass Moment of Inertia to quantify space-time patterns of urban growth

Wenwen Li, Elizabeth A. Wentz, Joanna Merson

School of Geographical Sciences and Urban Planning
Arizona State University
Tempe AZ 85287-5302
USA
Email: wenwen@asu.edu; wentz@asu.edu; jmerson@asu.edu

## 1. Introduction

The goal of this research is to demonstrate the capabilities of the Normalized Mass Moment of Inertia (NMMI) tool to quantify the spatiotemporal patterns of urban growth. Urban growth, or urbanization, is the transformation of the earth's land surface from predominantly unaltered surface materials into impervious surfaces, buildings, non-native plants, and other infrastructure. The positive implications of urbanization emphasize economic security, sense of community, and protection of outlying areas. The negative consequences of urbanization are associated with high instances of crime, large regions of poverty, high congestion, and poor environmental quality. A key feature to emphasizing the positive and diminishing the negative impact of urban growth in the face of a global population growth and climate change is through monitored information on the composition and configuration of population density and urban structures over time (Rosenzweig et al. 2011). Tools to quantify pattern provide insight into where and how cities are changing leading to better forecasting methods and models.

We utilize the NMMI pattern method to quantify change of city population over time. The NMMI is a pattern metric that considers both the geometric configuration of a spatial object (e.g., the urban boundary) as well as the attribute distribution within the object (e.g., population density). Quantifying both dimensions of pattern in a single normalized metric provides the means to evaluate change over time and the interaction between multiple variables. In this abstract we define the NMMI, describe its properties, and illustrate how it can be used to quantify space-time patterns of urban growth.

## 2. Normalized Mass Moment of Inertia (NMMI)

### 2.1 Definition

The premise of the NMMI is that pattern is influenced both by the geometric shape of an object as well as the density and distribution of attributes within the object. Before discussing the NMMI, we define its predecessor – the NMI (Normalized Moment of Inertia; Li et al. 2013) which measures the compactness of an areal object based purely upon its geometric shape. The compactness value can be obtained by comparing its area moment of inertia ($I_i^{G^i}$) to that ($I_0$) of a circle with the same area. Both area moment values are computed about an axis perpendicular to the shape's surface and passing through its centroid $G^i$. Known that a circle's area moment equals to its square area over two times π. The NMI value can be computed as:

$$C_{NMI} = \frac{I_0}{I_i^{G^i}} = \frac{A^2}{2\pi I_i^{G^i}} \quad (1)$$

This compactness value is based on the assumption that all elements on a shape are evenly distributed. To account for some property with uneven distribution pattern into compactness measure, Li et al. (2014) proposed a new approach, termed NMMI (Normalized Mass Moment of Inertia). Similar to Eq (1), the NMMI of a shape equals to the ratio between the mass moment of inertia and that of a circle with the same effective area.

$$C_{NMMI} = \frac{I_0'}{I_{i_{mass}}^{G_i}} = \frac{mA}{2\pi I_{i_{mass}}^{G_i}} \quad (2)$$

where m is the total mass, a.k.a. total quantity of some property in an area, A is the effective area, and $I_{i_{mass}}^{G_i}$ is the mass moment of inertia of a region, and $I_0'$ is the mass moment of the referencing circle.

## 2.2 Illustration of the method on urban growth

Urban growth results in different patterns based on the underlying conditions and localized properties such as proximity to rivers and mountain ranges, the location of transportation hubs and corridors, and prior population centers. To illustrate the potential of NMMI, assume we have two neighboring cities in which both are a square shape. We keep the urban boundary the same but we alter the distribution of growth. Toward the east, we have City A, with a density $\rho_1$ and toward the west we have City B with density $\rho_2$. The total larger region has a constant low density ($\rho_0 = 1$). Case I presents a scenario in which both cities' population density increase at the same rate, from very low – the same as $\rho_0$ to much higher ($\rho_1 = \rho_2 = \exp(30)$). Case II presents a second scenario that city A, as a developed city, has constant high population density ($\rho_1 = \exp(15)$) and city B's is newly developed city and is quickly attracting large immigration to a point that its population density is much higher ($\rho_1 = \exp(30)$) than city A. These illustrate possible different growth patterns in the two cities (Figure 1).

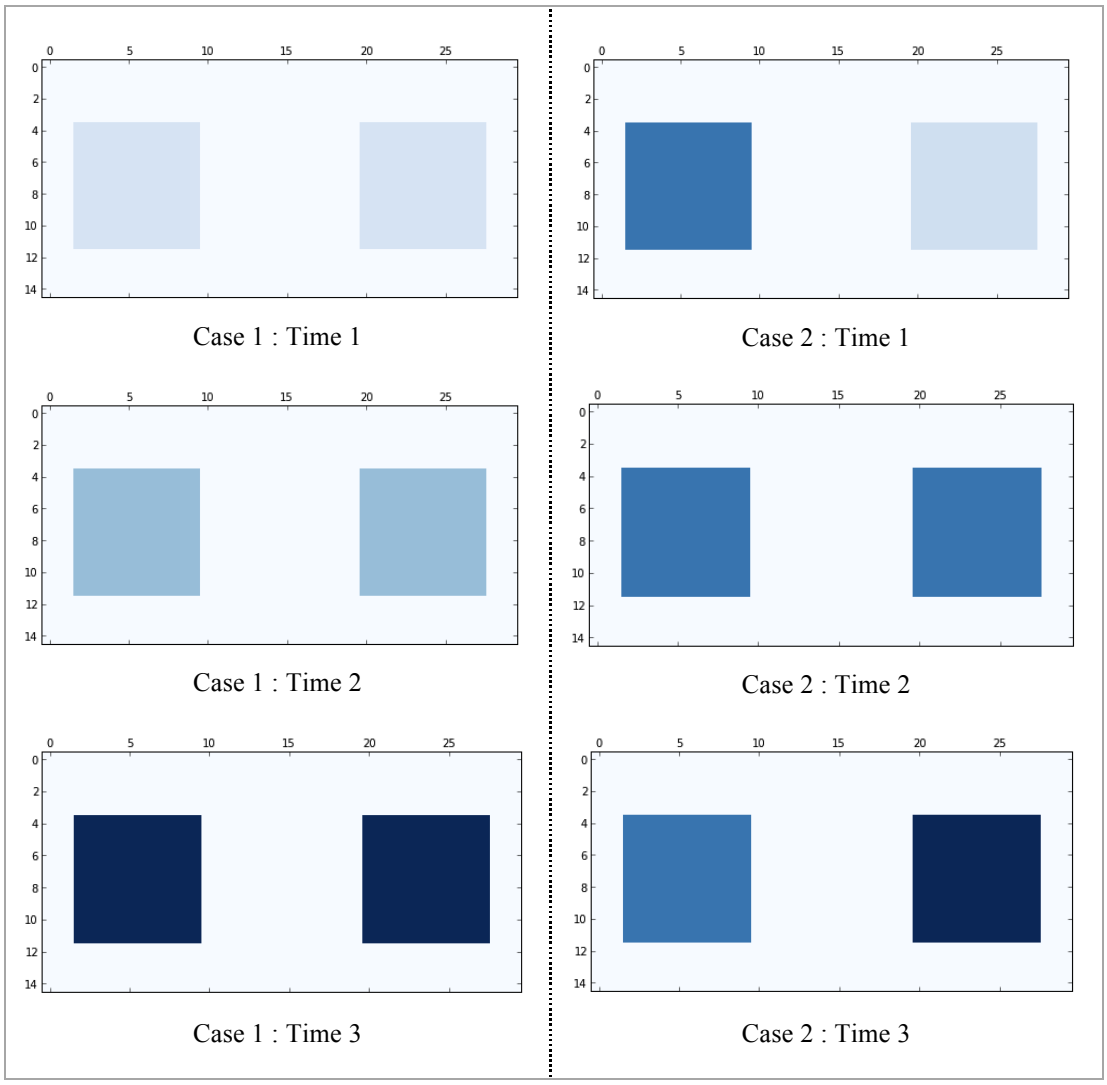



Figure 1. An urban growth scenario over space and time from Time 1 to Time 3. Case 1 shows growth of two neighbouring metro areas growing at the same rate. Case 2 shows the two neighbouring metro areas growing at two different rates.

To quantify the differences in pattern over time as shown in Figure 1, we apply the NMMI to the shape and density of these regions. Figure 2 shows two different methods to measure pattern over time. First we illustrate the results using a compactness index that only considers geometric boundary, such as NMI. Using NMI, the region will receive a constant compactness value 0.764 (grey line). Using this method, we do not have the

influence of the population density and therefore the pattern between Cases I and II are the same. Instead, we can measure change over time while considering population using a weighted compactness such as the NMMI. The compactness pattern shows a significant change when the population distribution changes inside of the region. In Case I, initially, both cities have the same density as the outsider regions, therefore, the population is equally distributed, so the NMMI yields the same value (0.764) as the NMI does. When both cities' population increases, the NMMI value starts to decline (the red curve), as population in this region starts to show dispersed pattern. When the two cities' density increase up to a certain value ($\rho_1 = \rho_2 = \exp(7)$ in this case), the NMMI reached the minimal value. At this point, the contribution of the part of region outside of both cities to the overall compactness can be neglected since the part's population density is very low in comparison to the two cities. The shape of the entire region can be considered to contain only the two cities. Later, the compactness curve becomes a straight line in parallel with the X axis, showing the continuing dominance effect of the two heavily populated cities to the overall compactness measure.

In Case II, a different scenario is presented: city A's density is constant ($\rho_1 = \exp(15)$) and is no longer growing, whereas city B is recently developed and started to grow in its population. The blue curve shows that initially, when city B's density is the same as the outside region, the overall compactness of this region obtained by NMMI is 0.955, which is the same as the shape compactness of the squared city A. Since city A has much more denser population than the surrounding area, its contribution to the compactness measure dominates, that is why the entire region's compactness becomes equal to that of city A. As city B develops, the population starts to scatter rather than concentrate within one city. Therefore, the compactness curve (blue line) starts to decline until when city B's density increases up to city A's density, when the population is most split and either city's shape take a major effect into the compactness measure. When city B's population density keeps increasing and is larger than city A, the compactness increases again until B's population is dominant in the entire region so the compactness value reaches its upper-bound, 0.955, the shape compactness value for city B.

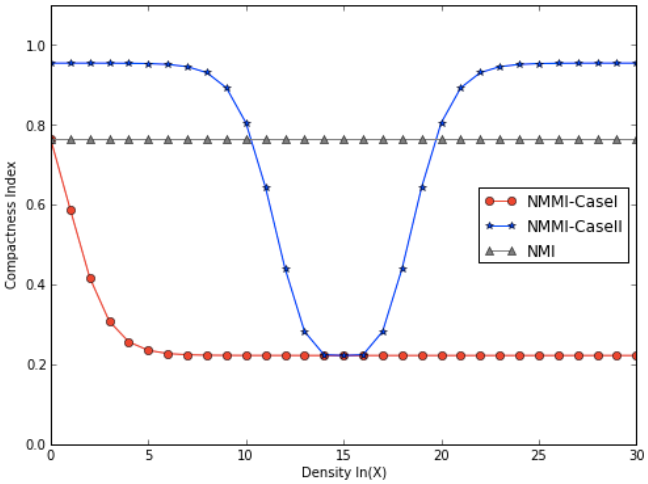



Figure 2. Compactness index of the region in each case

## 3. Concluding remarks

The results presented here illustrate how the NMMI tool can be applied to compare urban growth over time using a simulated dataset. We show how the distribution of a population across a region can be measured in conjunction with the underlying geometry. These results have implications on climate models as they begin to incorporate more variables associated with cities and their growth. Next step is to quantify real-world urban growth pattern using time-series data in Phoenix metropolitan area.

## Acknowledgement

This work is supported in part by the College of Liberal Arts and Sciences and the School of Geographical Sciences and Urban Planning Seed Grant Program at Arizona State University.

## References and Citations

Li W, Chen T, Wentz EA and Fan C, 2014, NMMI: A mass compactness measure for spatial pattern analysis of areal features. *Annals of Association of American Geographers*. (forthcoming)

Li W., Goodchild MF, Church RL, 2013, An efficient measure of compactness for 2D shapes and its application in regionalization problems. *International Journal of Geographic Information Science*, 27(6): 1227-1250.

Rosenzweig C, Solecki , WD, Hammer, SA, Mehrotra S, 2011, *Climate Change and Cities First Assessment Report of the Urban Climate Change Research Network* Cambridge University Press.

# Spatializing time in a history text corpus

A. Bruggmann and S. I. Fabrikant

University of Zurich, Department of Geography, Winterthurerstrasse 190, 8057 Zurich, Switzerland
Email: {andre.bruggmann; sara.fabrikant}@geo.uzh.ch

## 1. Introduction

Due to recent mass digitization initiatives of large text archives (e.g., GoogleBooks), the online access to massive amounts of text documents has risen dramatically. These efforts offer exciting new ways to explore human knowledge encapsulated in text. While text documents have been central to the humanities and the social sciences long before digitization, text sources are still largely untapped for spatio-temporal analyses in GIScience.

We aim to fill this gap and present a theory-driven framework that applies geographic information retrieval (GIR) and geovisual analytics (GeoVA) to an online dictionary about Swiss history. We chose the Historical Dictionary of Switzerland (HDS 2014) available in German, French and Italian as a prototypical online text archive, as it specifically includes spatial, temporal, and thematic information. Even though the 36,188 HDS documents implicitly contain spatio-temporal information, there are no such browsing or query possibilities in its current version. In own prior work (Bruggmann and Fabrikant 2014) we illustrate how spatial relationships between toponyms mentioned in text documents can be automatically extracted, re-organized semantically, and presented to an information seeker in static cartographic maps, and spatialized displays. In this paper, we specifically focus on how to automatically extract and visualize temporal information from the HDS, as to allow an information seeker to explore whether and how spatial relationships between historically relevant Swiss toponyms might have changed over time.

## 2. Methods

Following the methodology presented in Bruggmann and Fabrikant (2014), we first retrieved 169,094 toponyms from the HDS articles in German, by first identifying candidate toponyms with the Swissnames gazetteer (swisstopo 2014), and resolving disambiguation issues (Derungs and Purves 2014). We then re-organized the retrieved spatial data by assuming a (semantic) relationship between two toponyms, if they both co-occurred in the same article (Hecht and Raubal 2008). By example for this case study, we focus on the forty most often mentioned Swiss toponyms in the HDS. To analyze the potentially changing nature of toponym relationships over time, we employed *HeidelTime* (Strötgen and Gertz 2013) to automatically extract 510,357 temporal annotations from HDS text corpus, including *dates* (e.g., 07/09/1984), *periods of time* (e.g., 18$^{th}$ century) and *other temporal information*. In this paper, we exemplify our approach using centuries as the temporal unit of analysis, even though other temporal resolutions are possible. We used this temporal unit to weigh toponym relationships in each article. In other words, if two toponyms co-occur in articles that contain a high percentage of temporal annotations categorized as 20$^{th}$ century, their relationship is assigned a higher weight for the 20$^{th}$ century, compared to two toponyms that only co-occur in articles that have few annotations categorized as 20$^{th}$ century. Finally, we are able to visualize the extracted spatio-temporal toponym relationships, based on Fabrikant and Skupin's (2005) empirically validated spatialization framework.

## 3. Results and Discussion

We depict the extracted toponym relationships covering the last three centuries as a series of spatialized networks in Figure 1, where toponyms with stronger relationships are placed closer to one another on the network than those with weaker relationships. We constructed the network displays for each century separately, using the GEM layout algorithm to avoid edge crossings, and by applying the minimum spanning tree (MST) pathfinder algorithm available in the Network Workbench (NWB Team 2006) to visualize only the structurally most important relationships. Line width represents the strength of toponym relationships in the network. Toponym importance was calculated by summing all weighted relationships with all other toponyms in the network. Varying node sizes shows this: the larger the node, the higher the toponym importance in the network. We also ran the Blondel et al. (2008) community detection algorithm to investigate whether extracted toponym relationships might form node clusters that are more densely connected within the group, than with the rest of the network, and to identify whether these clusters might change over time. We visualized toponym clusters with differently colored nodes in Figure 1. Similarly, we depicted this information on a map of Switzerland, with the twenty most frequently occurring toponyms labeled for reference.

Focusing on the two most central nodes in the networks shown in Figure 1, i.e., *Zürich* (the financial capital) and *Bern* (the political capital), one can detect the steady increase of *Zürich*'s degree (i.e., the number of connected nodes) over time compared to *Bern*'s. While the degree for *Zürich* (14) and *Bern* (13) is about the same in the 19$^{th}$ century, *Zürich*'s degree rises to 15 nodes in the 21$^{st}$ century, compared to *Bern*'s, which drops to only eight. Hence *Zürich*'s well established importance as Switzerland's major economic hub today can be traced back with our semantic analysis of the HDS articles. Figure 1 shows that *Zürich*'s degree accelerated in the 20$^{th}$ and at the beginning of the 21$^{st}$ century.

Strikingly, Tobler's (1970) first law of geography ("Everything is related to everything else, but near things are more related than distant things") is also evident. The colored toponym nodes form contiguous spatial clusters in the maps in each time slice. The relationship dynamics of the blue and green clusters is interesting. The green toponym cluster *Sub North & West* appears in the 20$^{th}$ century as a sub-cluster of the blue colored *North & West* toponym cluster. One possible reason for this could be due to the separatist movements in the western parts of this region after WW II, resulting in the creation of the new Canton of *Jura* (located northwest of the city node labeled *Solothurn*) in 1979.

The central nodes *Zürich* and *Bern* are no longer located in the same cluster in the 21$^{st}$ century. *Zürich* now emerges as the center hub city for the eastern half of Switzerland, and *Bern* for the western half, respectively. The prior distinct toponym cluster in the Italian speaking region south of the Alps (i.e., *Lugano*, *Locarno*, *Bellinzona*) merges with the German speaking blue toponym cluster in the 21$^{st}$ century. One important reason for this, also connected to the rise of *Zürich*'s economic importance, may be the opening of the Gotthard road tunnel in 1980 which connects Southern Switzerland with its northern parts. The network visualizations provide another lens to view the hierarchical toponym relationship structure of over time, for example, by showing *Zürich*'s rising connectivity in the course of time, and also by detailing toponym hierarchies in hub nodes and peripheral nodes.

These encouraging results already illustrate how semantic analyses of space and time concepts extracted automatically from text documents in combination with geovisual analytics approaches can prove useful to assess the dynamics of spatial structure in a history text corpus over time.

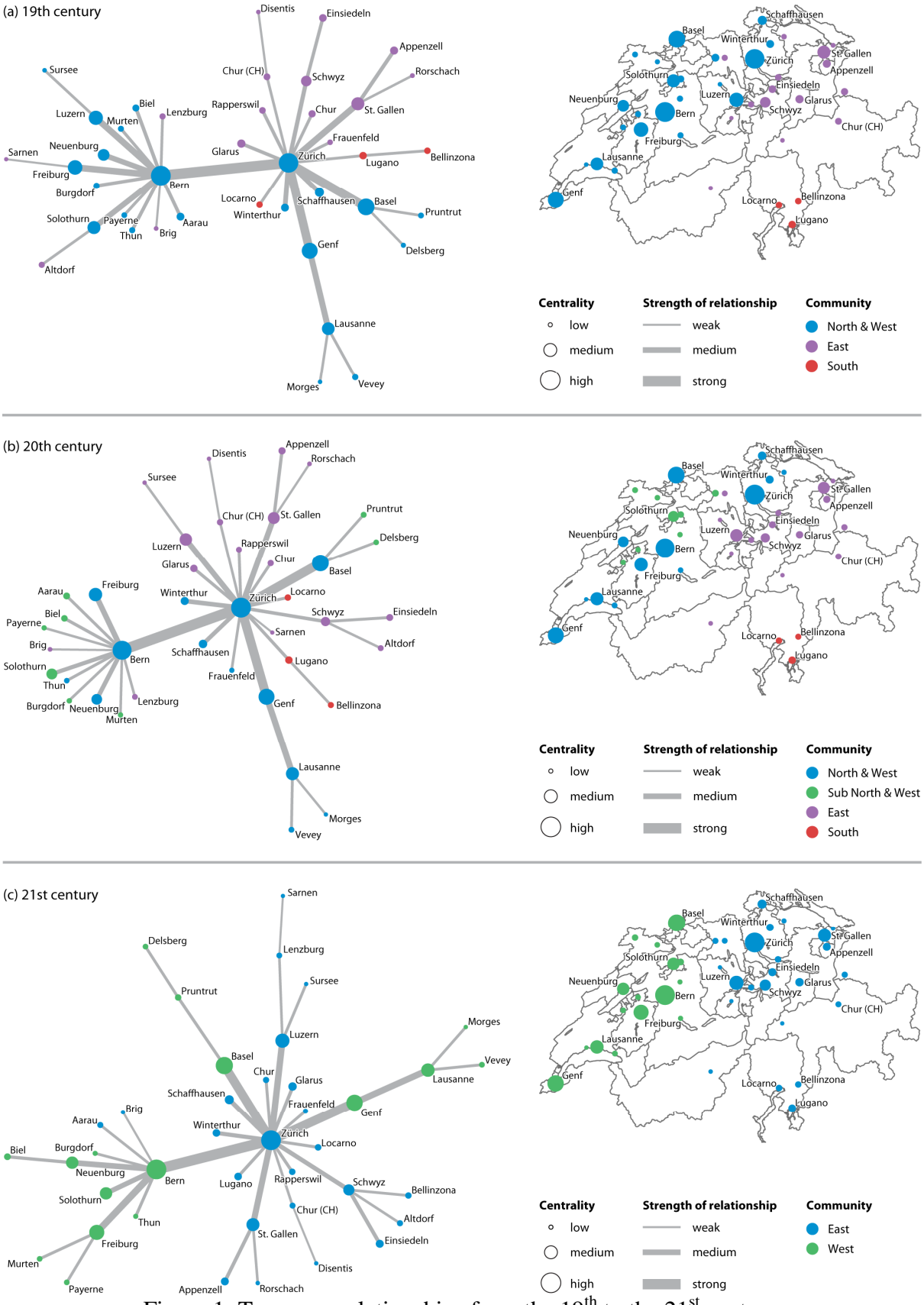



Figure 1: Toponym relationships from the 19th to the 21st century (map data source: swisstopo 2014).

## 4. Summary and Outlook

This paper introduces a novel text analysis framework based on GIR and GeoVA to automatically uncover and visualize latent spatio-temporal relationships buried in a history text corpus. In doing so, we hope to contribute our GIScience perspective to future interdisciplinary research projects in the digital humanities where space and time matter.

In future work, we aim to integrate thematic information analyses into our framework, as to identify the topicality of toponym relationships (e.g., economy, politics, culture, etc.) and how these might change over time. Finally, we will develop an interactive (online) user interface (e.g., using D3 technology) to extend the current HDS with spatio-temporal browsing and search capabilities.

## Acknowledgements

We would like to thank Curdin Derungs, Jannik Strötgen and Julian Zell who specifically helped us to implement the GIR part of our research. We are also grateful to Ross S. Purves and Damien Palacio for their invaluable feedback on this research project.

## References

Blondel V D, Guillaume J-L and Lambiotte R, 2008, Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. DOI: 10.1088/1742-5468/2008/10/P10008

Bruggmann A and Fabrikant S I, 2014, How to visualize the geography of Swiss history. In: Huerta, Schade, Granell (eds), Connecting a Digital Europe through Location and Place. *Proceedings,* International Conference on Geographic Information Science, AGILE 2014, Jun. 3-6, 2014, Castellón, Spain. ISBN: 978-90-816960-4-3.

Derungs C and Purves R S, 2014, From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science,* 28(6): 1272-1293. DOI: 10.1080/13658816.2013.772184

Fabrikant S I and Skupin A, 2005, Cognitively Plausible Information Visualization, In: Dykes J, MacEachren A M and Kraak M-J (eds), *Exploring Geovisualization*, 667-690.

Hecht B and Raubal M, 2008, GeoSR: Geographically Explore Semantic Relations in World Knowledge. In: Bernard L, Friis-Christensen and Pundt H (eds), *11th AGILE International Conference on Geographic Information Science*.

Historical Dictionary of Switzerland (HDS), 2014, http://www.hls-dhs-dss.ch/ (April 2014).

NWB Team, 2006, Network Workbench Tool 1.0.0. http://nwb.slis.indiana.edu (April 2014).

Strötgen J and Gertz M, 2013, Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2): 269-298.

swisstopo, 2014, SwissNames. http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html (April 2014).

Tobler W, 1970, A Computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2): 234-240.

# Correlating morphosyntactic dialect variation with geographic distance: Local beats global

Péter Jeszenszky, Robert Weibel

Department of Geography, University of Zurich (UZH),
Winterthurerstrasse 190, CH-8057, Zurich
Email: {peter.jeszenszky | robert.weibel}@geo.uzh.ch

## 1. Introduction

Similarly to Tobler's First Law of Geography, dialectology has its own postulate, termed the 'Fundamental Dialectological Postulate' (FDP): „Geographically proximate varieties tend to be more similar than distant ones" (Nerbonne & Kleiweg 2007: 154). This postulate seems intuitive, and thus several authors have tried confirming it by determining the degree of correlation between dialectal variation, expressed by a linguistic distance measure, and some geographic distance measure (e.g. Nerbonne & Kleiweg 2007; Spruit et al. 2009), all reporting (highly) significant correlations. While most authors have used Euclidean distance, some used travel time as a geographic distance measure that represents potential geographic language contact with an increased degree of realism (Gooskens 2004; Haynie 2012). However, in a recent study by Szmrecsanyi (2012) using corpus-based data about morphosyntax (i.e. grammatical constructs) in traditional English dialects, the FDP has been contested, reporting non-significant correlation.

The above studies are all rooted in linguistics, and have led to interesting results. From a geographical perspective, however, they all suffer from the crucial drawback of restricting the analysis to the —geographically speaking— global level, computing correlations for entire study areas, rather than exploring linguistic variation in more detail at the *local* level. Hence, they miss out on discovering regional differences in correlation structures, and on offering possible explanations of regionally different linguistic variation patterns. Also, global analysis alone will not be able to explain the large differences in the degrees of correlation reported in different studies.

Thus, the objective of our work is to enable the spatially differentiated comparison of linguistic variation and geographic distances, shedding new light on the FDP. For the case of morphosyntactic variation in Swiss German dialects, we present methods to establish global and local correlation between language and geographic distances, giving preliminary results and an outlook on possible extensions. While this work should be mainly beneficial for linguistics, we believe that it is also relevant to GIScience, since linguistic data represent a type of data that is uncommon in GIScience. Furthermore, we would like to show that dialectology and other strands of linguistics offer plenty of opportunities for GIScientists to contribute to advancing science at the interface between disciplines.

## 2. Data and Methods

### 2.1 Data

This study uses data from the Syntactic Atlas of German-speaking Switzerland (SADS; Bucheli & Glaser 2002). The SADS project was initiated in 2000 to map and study syntactical (i.e. grammatical) phenomena of Swiss German dialects. Close to 3,200 informants participated in a survey, providing answers to 118 questions, corresponding to

linguistic *variables*. Informants live in 383 municipalities, i.e. in approx. 25 % of the German speaking municipalities in Switzerland. An important feature of the SADS is that multiple informants occur per survey site, ranging between 3 and 26, with a median of 5 to 6 informants per site. Thus, linguistic variation, expressed by different *variants* for a given variable, exists also between respondents at each site. The following example shows this dual variation in a linguistic variable in the SADS:

English – 'I don't have enough change in order to buy a ticket.'
Standard German – 'Ich habe zu wenig Kleingeld um eine Fahrkarte zu lösen.'
Main variant 1. – 'Ich ha z wenig Münz **für** es Billet **z** lööse.'
Main variant 2. – 'Ich ha z wenig Münz **zum** es Billet **z** lööse.'

In this example, the linguistic *variable* is the syntax construct of the so-called infinitival complementizer, for which two *variants* exist, using 'für' and 'zum', respectively.

## 2.2 Methods

Linguistic (dis)similarity is often computed using edit distances, such as Hamming and Levenshtein distance (Spruit et al. 2009). However, since in the SADS multiple variants may occur per survey site, we had to use a different method. Figure 1, for two sample variables (Question I.01 and Question I.03) and two survey sites (Klosters, Flühli), shows the procedure of computing a linguistic distance — in this case, the *syntactic* distance — between a pair of sites.

Once the syntactic distances have been computed for all survey site pairs, the global correlation between the linguistic and the geographic distances between sites is computed. We use Pearson product-moment correlation and correlation established by the Mantel test.

Simply computing global correlations will not reveal the potential causes of linguistic variation, and is prone to ecological fallacy. This is improved in two ways. First, by focusing the analysis on a local subset of the study area. Second, by normalizing both the linguistic and geographic distances obtained, it becomes possible to compute residuals per site and thus analyze locally how well geographic distance predicts the observed linguistic distance.

Besides Euclidean distance, geographic distance was also represented by a travel time matrix provided by the Institute for Transport Planning and Systems at ETH Zurich (Fröhlich et al. 2004).

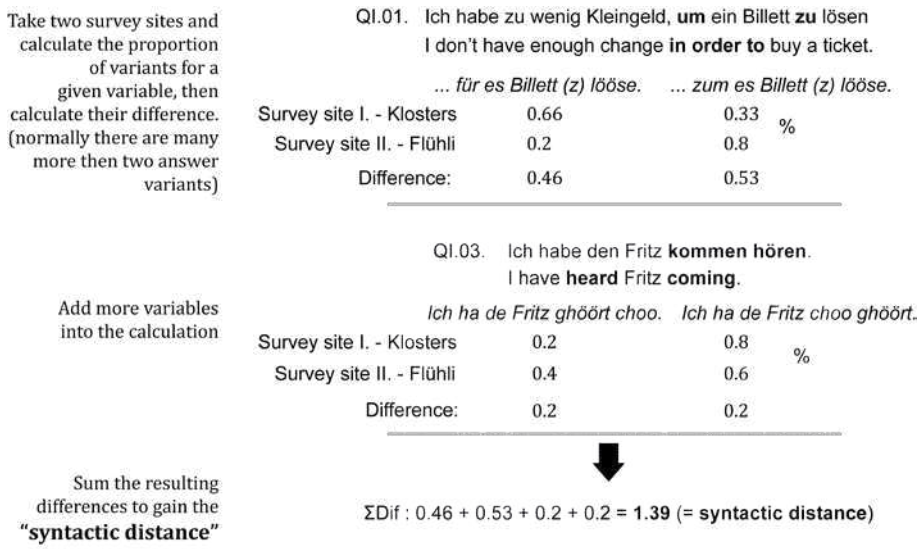



Figure 1: Workflow to compute the pairwise syntactic distance between two sites.

## 3. Results

So far, we have computed syntactic distances using 19 linguistic variables, which are hypothesized by the SADS linguists to be representative of the main morphosyntactic phenomena in Swiss German. Thus, the results reported below are preliminary from a dialectological perspective. However, they may nevertheless serve to illustrate the potential of our approach.

Tables 1 and 2 present the results of the correlation analysis on the global scale and for a particularly interesting local subset, the region between the Bernese Oberland and the German-speaking part of the Valais (BEOV, N = 45). All correlation coefficients are significant to highly significant (at least $p < 0.05$). As the right hand column of Tables 1 and 2 shows, the differences between the correlation coefficients at the global level as opposed to the coefficients at the BEOV level are significant, with the exception of the correlation the Mantel test results for both travel times. However, when comparing the correlations obtained with different distance measures, only very few were reported significant (results not shown in Tables 1 and 2). Only subtle differences between 0.722 and 0.747 exist for the global level and are thus not significant. In the BEOV subset, only one highly significant difference ($p < 0.01$) can be found between Euclidean distance and travel times 1950 in the Mantel test (0.366 vs. 0.750). Additionally, the difference between Euclidean distance and travel times 2000 in the Mantel test (0.366 vs. 0.707) is significant ($p < 0.05$). And one difference—between Euclidean distance and travel times 2000 in the Pearson correlation coefficients (0.307 vs. 0.578)—is almost significant ($p = 0.0582$).

The map in Figure 2 shows the survey sites, represented as Voronoi polygons to fill in the gaps between sites, colored according to their syntactic distance from a particular place, Schaffhausen, with the borders of the Swiss cantons overlaid. Normalizing the distances, residuals per site can be obtained, showing the degree of agreement between the two distance measures (Fig. 3). Thus, if the normalized syntactic distance from the survey site "Obersaxen" were in perfect linear agreement with the corresponding normalized Euclidean distance, no residuals would show in Figure 3. Figure 4 then maps the residuals of Figure 3 to geographic space. Finally, Figure 5 depicts the syntactic distances from "Adelboden" for the local subset BEOV in the area of the Bernese Oberland and the German speaking part of the Canton of Valais.

Table 1. Pearson correlation coefficients for global area and a regional subset.

| For 19 variables | Syntactic distance (global, N = 383) | Syntactic distance (BEOV subset, N = 45) | Fisher's Z, one-tailed |
|---|---|---|---|
| Euclidean distance | $0.722^{***}$ | $0.307^{*}$ | *** |
| Travel times by car - 1950 | $0.745^{***}$ | $0.578^{***}$ | * |
| Travel times by car - 2000 | $0.743^{***}$ | $0.524^{***}$ | * |

* = P ≤ 0.05, ** = P ≤ 0.01, *** = P ≤ 0.001, ns = statistically not significant

Table 2. Mantel test results for global area and a regional subset.

| For 19 variables | Syntactic distance (global, N = 383) | Syntactic distance (BEOV subset, N = 45) | Fisher's Z, one-tailed |
|---|---|---|---|
| Euclidean distance | $0.747^{***}$ | $0.366^{**}$ | *** |
| Travel times by car - 1950 | $0.738^{***}$ | $0.750^{***}$ | ns |
| Travel times by car - 2000 | $0.734^{***}$ | $0.707^{***}$ | ns |

* = P ≤ 0.05, ** = P ≤ 0.01, *** = P ≤ 0.001, ns = statistically not significant

## 4. Discussion

As Tables 1 and 2 show, all correlation coefficients are highly significant on the *global level*, independently of the correlation measure used. However, the difference between the results for the different geographic distance measures is not statistically significant.

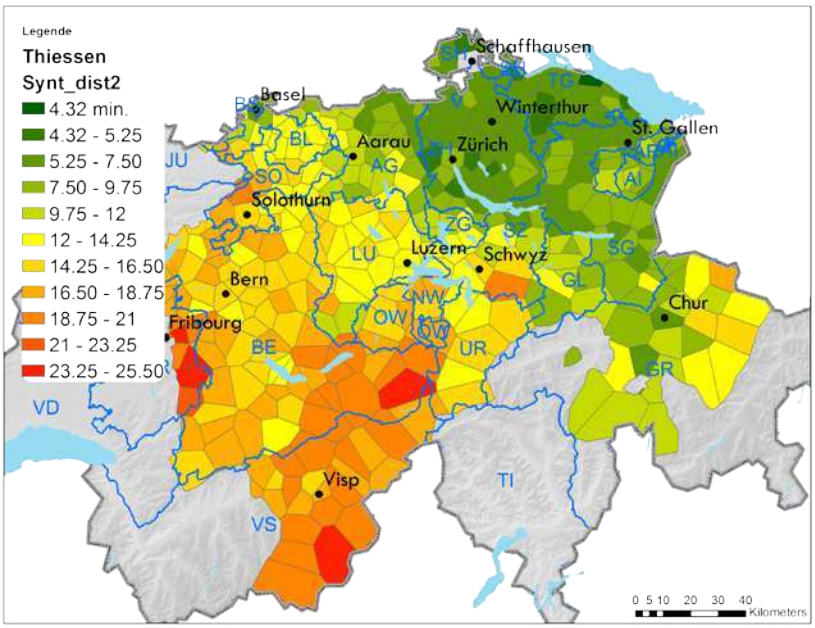



Figure 2: Syntactic distances from Schaffhausen.

The story is different at the *regional level*, represented by the BEOV subset. Here, we find generally lower correlations compared to the corresponding values at the global level, but we also find significant differences between the Euclidean and travel time distances. In the BEOV subset, a high mountain area is represented, where topography crucially influences travel times. Thus, travel time is a significantly better predictor at this more local level.

As Figure 2 shows for the example of Schaffhausen, the syntactic distances from this site exhibit a pattern that appears to largely follow the increase in Euclidean distance, with some exceptions. This suggests a possible explanation of the highly significant correlation with Euclidean distance on the global level, which at the same time does not differ significantly from correlation results obtained with travel times.

The differences between normalized syntactic and Euclidean distances (Fig. 3) follow a decreasing trend. They are positive at short ranges, meaning that the Euclidean distance underestimates short-range syntactic variation. The opposite is the case at long ranges, where Euclidean distance overestimates syntactic variation. This overestimation at long ranges makes sense, since geographic distance increases continuously, while the dialectal distance may only increase to a certain level. If two dialects become too dissimilar, they will be considered two different *languages*, as they are no longer mutually intelligible. This geographic pattern becomes even more apparent in the map of Figure 4.

Finally, Figure 5 shows some interesting patterns at the regional and local level for the BEOV subset, which represents high mountain topography, with secluded valleys. These patterns would not become apparent if the analysis was restricted to the global level. For instance, we could see a bridging effect of two mountain passes, the Gemmi Pass and the Grimsel Pass, respectively, which connect two sides of a high mountain range that largely exceeds 4,000 m.a.s.l. The Gemmi Pass being one of them, nowadays cannot be traversed by road but used to be a major pass in the Middle Ages when most dialect formation took place. Further work, however, is needed to explore these effects in more detail.

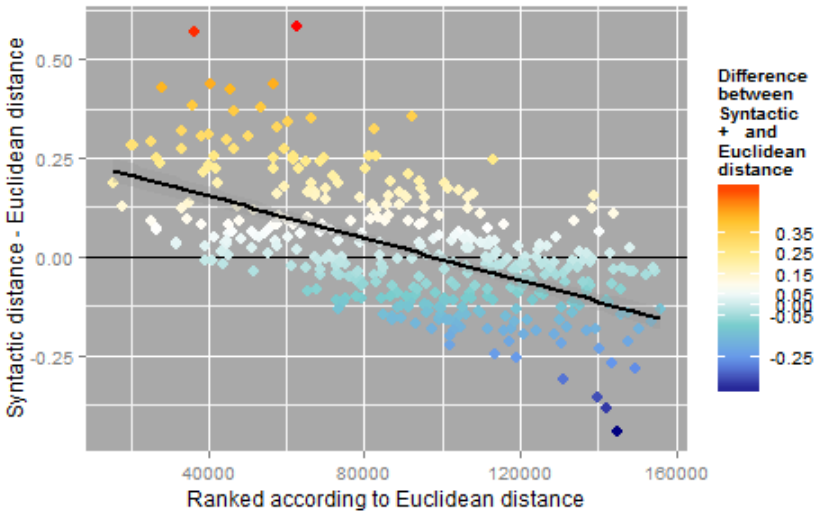



Figure 3: Residuals of syntactic distance and Euclidean distance for survey sites paired with the alpine village Obersaxen (cf. Figure 4).

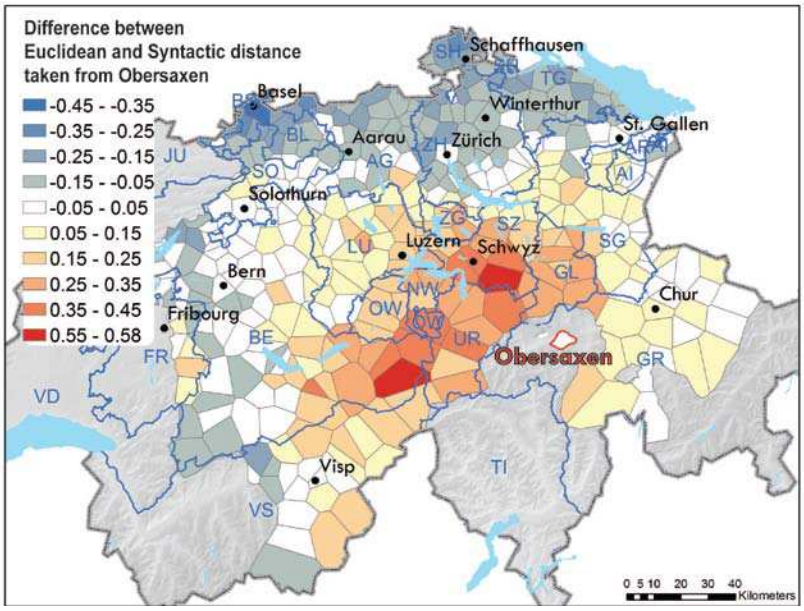



Figure 4: The residuals of Figure 3 mapped to geographic space.

## 5. Conclusions

We have shown how global correlation analysis with geographic distances in dialectology can be extended to the local level, painting a more differentiated picture of the dialectal variation across space. For the case of morphosyntactic variation represented by the SADS, we have been able to confirm the FDP, and show that different geographic distance measures only play out at the local level as a predictor variable.

Various extensions are possible. From a linguistic perspective, we will add more SADS variables and possibly also variables from other linguistic levels (lexis, phonetics, morphology). While today, travel times are increasingly approximating the concentric pattern of Euclidean distance, owing to ever improving accessibility, we will be extending the analysis to pre-1850 travel times, hypothesizing the results to differ significantly from those obtained with Euclidean distances. We will also explore other proxies of language contact such as linguistic gravity (Szmrecsanyi 2012), commuter matrices etc.

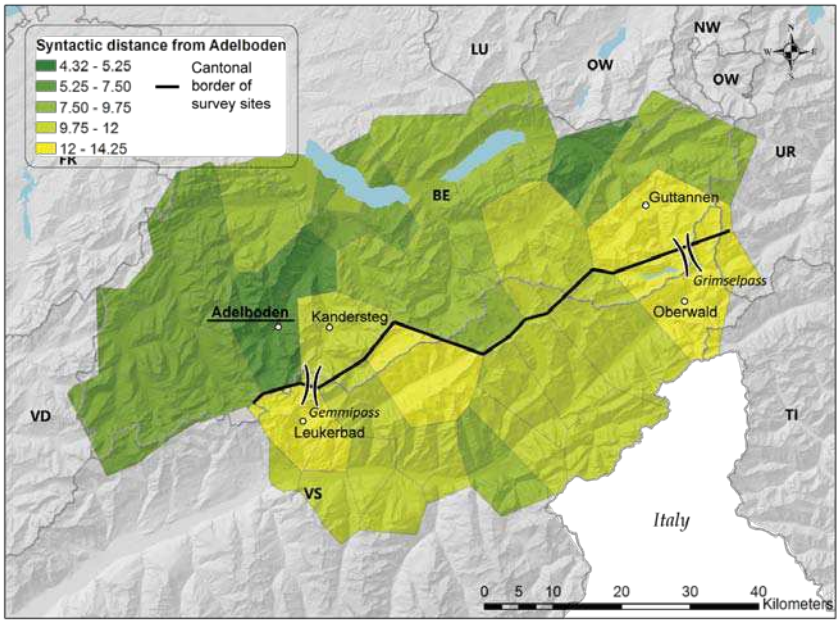



Figure 5: Map of syntactic distances from Adelboden in the BEOV subset. The cantonal border is formed by a major alpine drainage divide, bridged by two mountain passes.

From the methodological perspective, the current method of linear summation of syntactic distance assumes independence of variables, neglecting potential mutual correlation. Correlation analysis and dimension reduction could be explored. Finally, the most interesting extension will be to represent "geography" not only by geographic distances, but attempt to relate linguistic (i.e. syntactic) variation to geographical features, such as topographic, political or cultural borders.

## Acknowledgements

This research represents part of the PhD project of the first author. Funding by the Swiss National Science Foundation through project SynMod (CR12I1-140716) is gratefully acknowledged. We are grateful to the Institute for Transport Planning and Systems within the Swiss Federal Institute of Technology for providing the travel time data, and for the Syntactic Atlas of German-speaking Switzerland (SADS) project for the syntactic data. Finally, we would like to thank Philipp Stöckle, German Department of UZH, for his valuable comments.

## References

Bucheli, C., & Glaser, E. (2002). The Syntactic Atlas of Swiss German Dialects: Empirical and Methodological Problems. In S. Barbiers, L. Cornips, & S. van der Kleij (Eds.), *Syntactic Microvariation* (Vol. 2., pp. 41–73). Amsterdam: Meertens Institute Electronic Publications in Linguistics.

Fröhlich, P., Frey, T., Reubi, S., & Schiedt, H. U. (2004). *Entwicklung des Transitverkehrs Systems und deren Auswirkung auf die* Raumnutzung *in der Schweiz (COST 340): Verkehrsnetz-Datenbank* (No. 340) (p. 54).

Gooskens, C. (2004). Norwegian Dialect Distances Geographically Explained. In *Language Variation in Europe. Papers* from *the Second International Conference on Language Variation in Europe ICLAVE Vol. 2. 2004.* (p. 10).

Haynie, H. J. (2012). Studies *in the History and Geography of California Languages*. University of California, Berkeley.

Nerbonne, J., & Kleiweg, P. (2007). Toward a Dialectological Yardstick. *Journal of Quant. Ling.*, *14*(2), 148 p.

Spruit, M.R., Heeringa, W. & Nerbonne, J. (2009). Associations among Linguistic Levels. *Lingua*, 119(11), 1624–1642

Szmrecsanyi, B. (2012). Geography is overrated. In S. Hansen, C. Schwarz, P. Stoeckle, & T. Streck (Eds.), *Dialectological and Folk Dialectological Concepts of Space* (pp. 215–232). Berlin, Boston: De Gruyter.

# Why landscape terms matter for mapping: A comparison of ethnogeographic categories and scientific classification

Flurina M. Wartmann[1] & Ross S. Purves[1]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
Email: {flurina.wartmann; ross.purves}@geo.uzh.ch

## 1. Introduction

Categories are central in the way we structure information about the world around us and form the basis for representations in GIS. However, the translation of natural language concepts and categories into formal GIS environments is complicated by the fact that different terms can be used for the same phenomenon or the same terms can be understood in different ways (Harvey et al. 1999, Bishr 1998). Semantic interoperability remains a challenge even where it applies to seemingly straightforward terms such as 'forest', as conceptualizations of the phenomenon vary between different communities of practice, resulting in different classifications (Comber et al. 2005) with implications for management of these areas (Robbins 2001).

Ontologies as specifications of certain conceptualizations are important for developing formalized representations in GIS (Schuurman 2006). However, in building an information system, the question is 'where to take the ontology from'. One approach is to use scientific classifications, which has been criticized for imposing conceptualizations that fail to take into account how local people perceive, refer to and interact with landscape (Rundstrom 1995). Given the importance of GIS in spatial planning and natural resource management, there is a need to consider how to better elicit and represent such local concepts and categories and how multiple competing ontologies can be represented (Turnbull 2007).

In this respect, folk categories can provide the basis for ontology development (Wellen and Sieber 2013, Kuhn 2001, Smith and Mark 2001). The field of ethnophysiography, positioned between GIScience, social anthropology and linguistics deals with folk categorizations of the geographic domain, focusing on how different speech communities refer to and categorize landscape features including landforms and vegetation assemblages, as well as the cultural beliefs and customs related to those features (Mark et al. 2011, Mark and Turk 2003). Here, we present initial findings on the comparison of ethnogeographic categories with a scientific classification in the Bolivian Amazon.

## 2. Methods

As is common in ethnophysiography, we adopted a set of ethnographic methods including field walks and semi-structured interviews on landscape pictures to elicit terms for geographic features. We conducted our study in the Madidi National Park, established in 1995 to protect the region's high biological and cultural diversity. In the study area along the Beni river, people self-identify as Takana, an indigenous group with about 5,000 people, of whom the majority are now Spanish monolingual speakers. Contemporary Takanan lifestyles are based on a mixture of hunting, fishing, subsistence agriculture and wage-labour.

We collected data for this study over a period of 7 months from 2012 to 2013, with a total of 14 interviews held in Spanish.

## 3. Results

We documented 158 generic Spanish terms for geographic features. The most terms are coined for vegetation units, followed by those related to agriculture, water and topography. In the following, we focus on vegetation as an integral part of the landscape (and not simply land cover or land use) covering most of the land surface in our study area.

Out of 59 identified vegetation related landscape units, most are named after plants that have specific local uses. One example is the term *balsal* for an area that consists of *balsa* trees (*Ochroma pyramidale*). The Takana use a *balsal* as an area where they harvest balsa trees for building rafts and cut off bark to use as ropes. This example illustrates how most of the local landscape terms are monolexical and linguistically transparent. By adding the Spanish suffix '*-al*' to a plant name, it becomes a generic landscape term.

The 59 local terms for vegetation units differs from an existing botanical classification with 15 broad vegetation units (Fuentes 2005). More importantly, we also observed differences at a more conceptual level. Certain terms such as *monte alto* ('forest', Table 1) are spiritually significant, as they are believed to be inhabited by forest spirits, where certain rules need to be followed when entering or extracting resource in such areas.

Table 1. Examples from local terms and a scientific botanical classification

| Local term | Scientific classification |
|---|---|
| *balsal* | Riverine vegetation characterized by *Ochroma pyramidale* |
| *barbecho* | Lowland Amazonian forest |
| *charral* | Pioneer riverine scrub vegetation on sandy soils characterized by *Gynerium sagitattum* |
| *japainal* | Seasonally flooded Amazonian forest characterized by *Heliconia episcopalis* |
| *monte alto* | Lowland Amazonian forest |

An apparent mismatch between the local and the scientific conceptualization is illustrated through the landscape term *barbecho* (Table 1). For the Takana, a *barbecho* is an old agricultural field left fallow that can be re-planted again. However, due the dense herbal layer and tall trees used as border markers by the Takana, the National Park administration classified these forest patches as 'primary rainforest', leading to exclusion of local people.

## 4. Discussion and Conclusion

We have shown that the ethnogeographical categories of the Takana in Bolivia consist of at least 158 terms, with most terms being coined for vegetation units. As these terms are commonly used in direct speech and are linguistically simple, they can be seen as 'basic terms' (Tversky and Hemenway 1984). These 'folk generic terms' are more diversified than the scientific classification and provide valuable information for developing more appropriate classification systems in which the spatial categories to be represented in a GIS can be locally grounded (Wellen and Sieber 2013, Mark et al. 2011). However, this local grounding then also needs to be translated into more culturally appropriate GIS, which takes into account the varied local understandings of landscape. Such understandings are intimately connected to the environment and specific livelihoods of a speech-community. In the arid lands of Australia, the Yindjibarndi for instance have a diversified vocabulary for hydrological features that contain the magnitude of water flow (Turk et al. 2011), while the Gitskan in

Canada distinguish different snowfields, avalanche tracks and cliffs that reflect their need for a vocabulary describing travel routes and hunting areas in mountainous terrain (Johnson 2011).

Such folk classifications and differences with formal scientific classifications are not merely local curiosities, but have consequences for how these areas are classified and ultimately managed. Given the importance of GIS in landscape planning and management, the need remains to consider how to more adequately represent multiple ontologies (Turnbull 2007).

## Acknowledgements

The 'Consejo Indigena del Pueblo Takana' and the National Park Authorities (SERNAP) granted research permits. We acknowledge funding by the 'Forschungskredit' of the University of Zurich, grant no. FK-13-104 and financial support for fieldwork from Hans Vontobel, Maya Behn-Eschenburg, Ormella and Parrotia foundation.

## References

Bishr, Y, 1998, Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science* 12 (4): 299–314.

Comber, A, Fisher, P and Wadsworth, R, 2005, What is land cover? *Environment and Planning B: Planning and Design* 32 (2): 199–209.

Fuentes, A, 2005, Una introducción a la vegetación de la región de Madidi. *Ecología en Bolivia* 40 (3): 1–31.

Harvey, F, Kuhn, W, Pundt, H, Bishr, Y and Riedemann, C, 1999, Semantic interoperability: a central issue for sharing geographic information. *The Annals of Regional Science* 33 (2): 213–232.

Johnson, LM, 2011, Language, landscape and ethnoecology, reflections from northwestern Canada. In Mark, DM, Turk, AG, Burenhult, N and Stea, D (eds), *Landscape in language. Transdisciplinary perspectives*. Amsterdam / Philadelphia, John Benjamins Publishing, 291–326.

Kuhn, W, 2001, Ontologies in support of activities in geographical space. *International Journal of Geographical Information Science* 15 (7): 613–631.

Mark, DM, Turk, AG, Burenhult, N and Stea, D (eds), 2011, *Landscape in language. Transdisciplinary perspectives*. Amsterdam / Philadelphia, John Benjamins Publishing.

Mark, DM, and Turk, AG, 2003, Landscape categories in Yindjibarndi: ontology, environment and language. In Kuhn, W, Worboys, M and Timpf, S (eds), *Spatial Information Theory. Foundations of Geographic Information Science*. Lecture notes in computer science, 2825, Berlin, Springer, 28–45.

Robbins, P, 2001, Fixed categories in a portable landscape: the causes and consequences of land-cover categorization. *Environment and Planning A* 33 (1): 161–179.

Rundstrom, RA, 1995, GIS, Indigenous peoples, and epistemological diversity. *Cartography and Geographic Information Systems* 22 (1): 45–57.

Schuurman, N, 2006, Formalization matters: critical GIS and ontology research. *Annals of the Association of American Geographers* 96 (4): 726–739.

Smith, B; Mark, DM, 2001, Geographical categories: an ontological investigation. *International Journal of Geographical Information Science* 15 (7): 591–612.

Turk, AG, Mark, DM and Stea, D, 2011, Ethnophysiography. In Mark, DM, Turk, AG, Burenhult, N and Stea, D (eds), *Landscape in language. Transdisciplinary perspectives*. Amsterdam / Philadelphia, John Benjamins Publishing, 25–45.

Turnbull, D, 2007, Maps narratives and trails: performativity, hodology and distributed knowledges in complex adaptive systems - an approach to emergent mapping. *Geographical Research* 45 (2): 140–149.

Tversky, B and Hemenway, K, 1984, Objects, parts, and categories. *Journal of Experimental Psychology: General* 113 (2): 169.

Wellen, C and Sieber, R, 2013, Toward an inclusive semantic interoperability: the case of Cree hydrographic features. *International Journal of Geographical Information Science* 27 (1): 168–191.

# Temporal Analysis of Georeferenced Emotions Extracted From Photo Metadata

D. Burghardt[1], A. Körner[1], E. Hauthal[1]

[1]Dresden University of Technology, Institute of Cartography, 01062 Dresden
Email: {dirk.burghardt; eva.hauthal}@tu-dresden.de, andreas.koerner87@web.de

## 1. Introduction

Current location based services mainly provide objective information and collections of facts. Subjective components such as emotions and opinions can provide additional alternative information useful in decision making, e.g. in tourism, business, entertainment and the like. Therefore research on affect analysis was carried out by capturing and analysing georeferenced emotions from user-generated content (UGC). An approach was developed for extracting location-based emotions from the written language in the metadata of georeferenced Flickr and Panoramio photos. These data describe places and thus contain the sense of users as places and emotions are connected fundamentally. The approach was applied to the study area of Dresden, Germany.

## 2. Data Derivation

For gathering emotional data, emotions need to be structured. In psychology different approaches for structuring emotions exist and can be distinguished into dimensional and differential approaches (Schimmack 1999). Differential approaches emphasize the distinguishable subjectively experienced qualities of emotions (Izard 1977). Dimensional approaches try to reduce affective states to a few dimensions. Thus each emotion can be described as a combination of different severities of those dimensions. We are working with the model proposed by Russell (1980) involving the two dimensions valence and arousal which can be described as ranging from positive/pleasing to negative/displeasing and from arousing/intense to unarousing/numbing (see Figure 1). With the help of these two dimensions it is possible to locate emotions within valence-arousal-space. For instance joy is a very positive emotion with high arousal whereas anger also has a high arousal but a negative valence.

Our approach applies several methods of natural language processing to words that are contained in the title, description and tags of georeferenced Flickr and Panoramio photos (Hauthal and Burghardt 2013). All words are matched with two emotional word lists: ANEW (Affective Norms for English Words; Bradley and Lang 2010) and BAWL-R (Berlin Affective Word List Reloaded; Võ et al. 2009) comprise words that are weighted with a valence and an arousal value and reflect affective connotations (Hayakawa, 1952) and thus represent emotions. ANEW covers 2476 English words, while BAWL-R contains 2901 German words. Each word is stored together with its emotional values from ANEW or BAWL-R and with the coordinates of the respective photo. Within this extraction approach, various grammatical issues were considered, like negations of words or amplifications. Procedures were developed for modifying the emotional values of the affected word, for example for inverting or intensifying them.

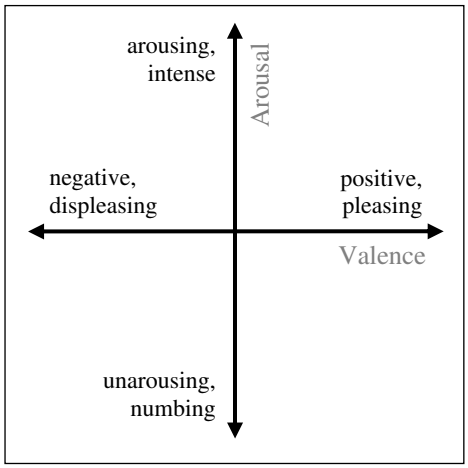



Figure 1. Two-dimensional structure of emotions by Russell (1980).

We applied the algorithm for extracting georeferenced emotions from photo metadata to a dataset of 52,954 Flickr and Panoramio photos of Dresden from 4,344 users covering a period starting at the launch date of Flickr (February 2004) and Panoramio (October 2005) until 2013-07-05. The data were requested with the respective REST API. The photo metadata used altogether contained 792,089 words. 116,780 of those words or the respective synonyms/hypernyms of them were found in the emotional word lists ANEW and BAWL-R.

## 3. Temporal Aspects

While analysing the emotional data we realised that one place is not necessarily connected with only one emotion. This can have two reasons. The first argument is based on the consideration of personal preferences, experiences or memories. For example a very scenic park might be admired by most of the people but if someone remembers that their boyfriend or girlfriend broke up with them in this park, then this person probably does not like the park anymore because of this personal experience. This shows that individual and collective emotions need to be distinguished. The second reason could be temporal aspects. Some years or even decades ago a place might have evoked different emotions than it does currently but these former emotions can still be detected with the help of Flickr and Panoramio photos. With regard to this, it seems obvious to investigate the time-dependency of emotions: are there places that are more attractive in the summer or in the winter time, in the day time or in the night? However for this paper we focus on three certain kinds of temporal aspects: long-term trends, periodic events and single events.

## 4. Application Examples

### 4.1 Long-term Trends

Since the used dataset covers several years, it can be analysed regarding recurrent frequencies of photos over the course of a year and whether the photo frequency correlates with the number of emotions. For this examination, photo data for the years 2009, 2010, 2011 and 2012 were analysed (see Figure 2); the years before were disregarded since only few photos were taken. The emotional data were divided into four quadrants based on the structure of valence-arousal-space (compare Figure 1). The four resulting quadrants are combinations of positive/negative valence and high/low arousal.

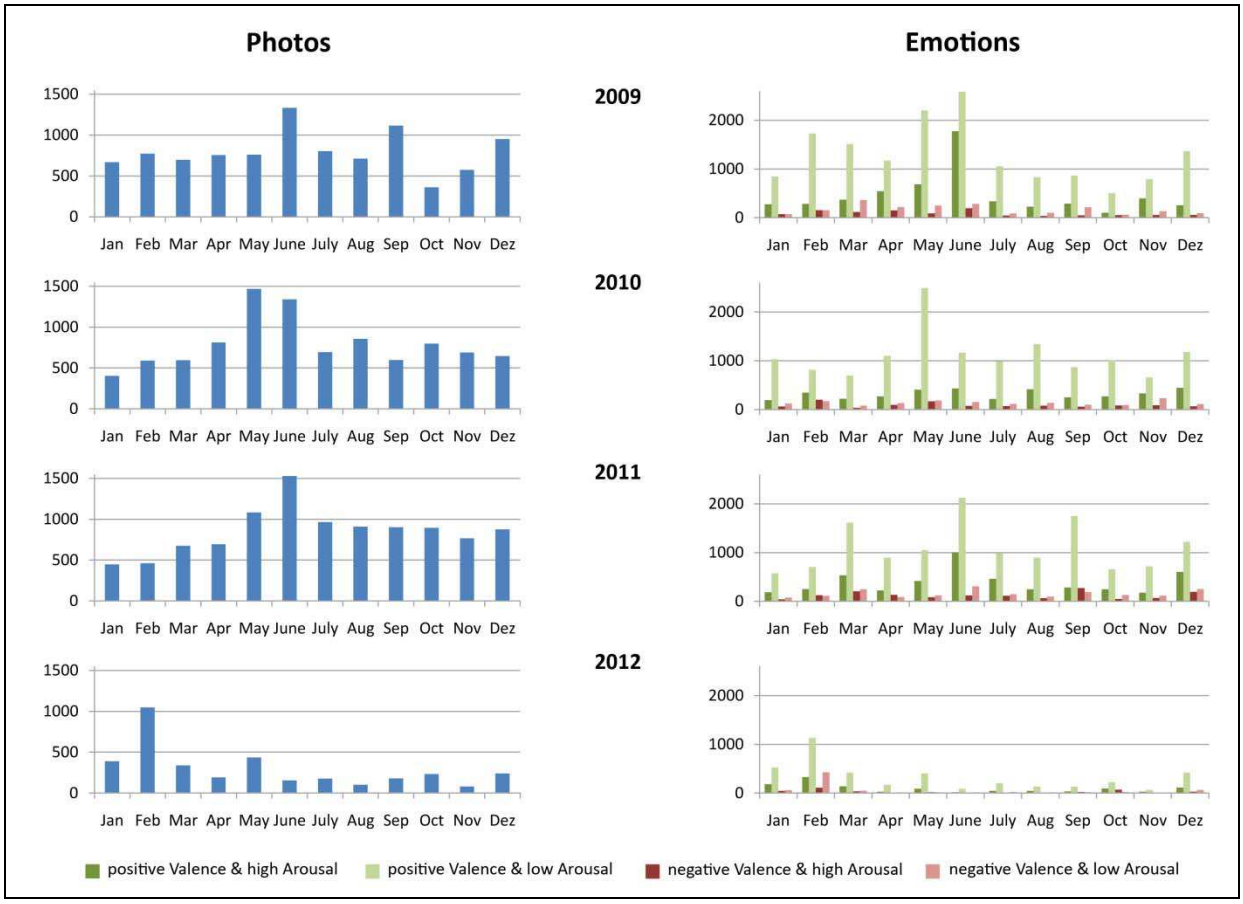



Figure 2. Temporal distribution of photo and emotion frequency over the course of the years 2009, 2010, 2011 and 2012.

The left part of Figure 2 shows the distribution of photo quantity for all four years while the right part shows the distribution of emotions occurring during these years divided into the four quadrants. The temporal distribution of photos over one year is similar for 2009, 2010 and 2011 and can be explained by tourist activities. The number of emotions corresponds with the number of photos: when more pictures were taken, more emotions could be detected. In general, more positive than negative emotions were extracted, especially positive emotions of low arousal.

## 4.2 Periodic Events

On the 13th of February 1945 huge parts of Dresden were destroyed by allied air attack. Each year a remembrance of this bombing takes place on one or two days in February. For the past 15 years more and more right-wing extremists use this event for their own propaganda purposes. As a reaction to that, counterdemonstrations have been organised and in the most recent years there have been confrontations and riots on both sides.

For recognising this event in the emotional data of Dresden, the years 2006 to 2013 were investigated. Two days prior to and past the demonstration date of each year are also considered (see Figure 3). The days of demonstrations are printed bold. It is noticeable that on the days of demonstrations more emotions were detected than on the two days before and after. The emotions on these days are mainly negatively arousing. These events are described by words with negative connotations like ‘police’, ‘Nazi’, ‘attack’ or ‘damage’. They are an extraordinary, although recurrent occurrence: usually nothing negatively arousing can be found at this place except on those one or two days in February.

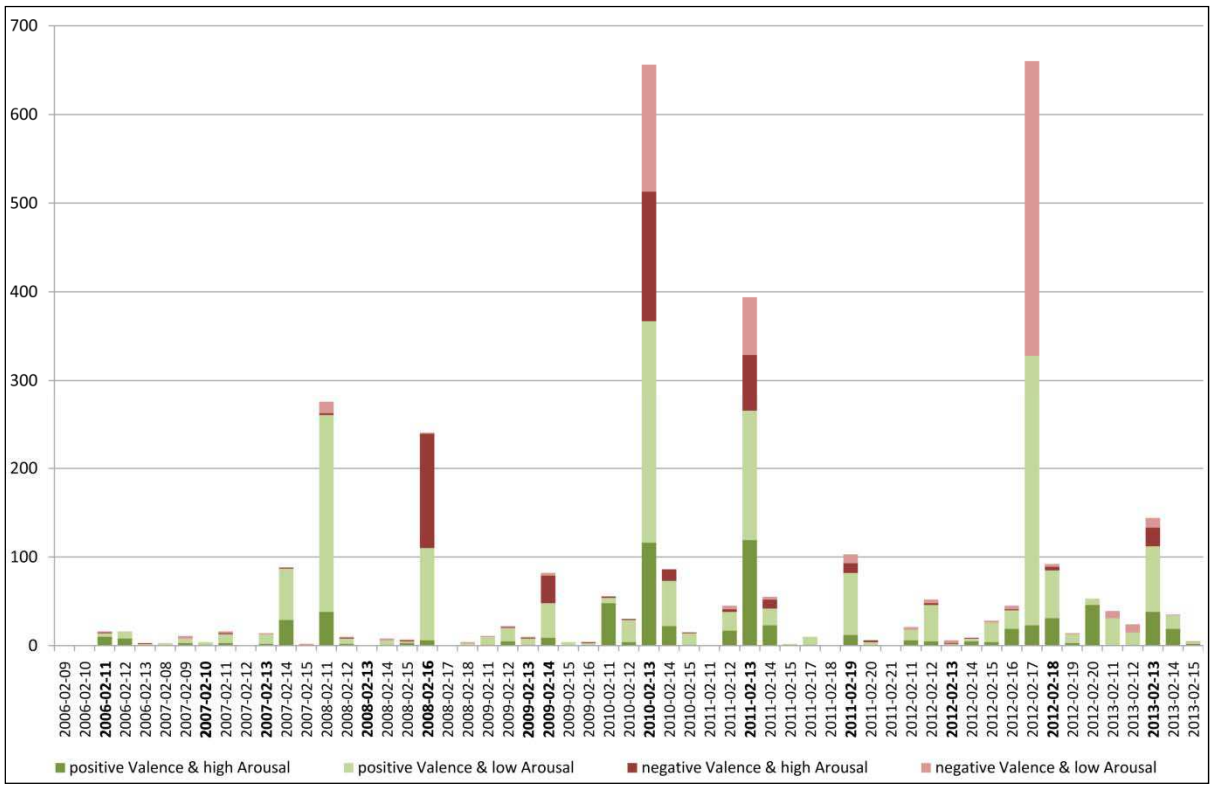



Figure 3. Temporal distribution of emotion frequency for the time of February demonstrations.

### 4.3 Single Events

Single events are happenings which occur only once a time and can be analysed as already known events or can be detected by analysing emotional peaks in the data. For this work the single event of the Elbe River Flood in June 2013 has been studied. The Elbe River has usually a water level of about 2 metres, but in the beginning of June 2013 it climbed to 8.76 metres within a few days.

Considering the entire month, in the 11 days of the flood (2013-06-02 until 2013-06-12) 66% of all photos were taken and thus a correspondingly large number of emotions (67%) was detected for these days. This reveals that single events, their temporal extent and their emotional characteristics can be detected by the number of photos. The emotional peak regarding the flood can be identified for the 4$^{th}$ June (see Figure 4), probably the day with highest uncertainty, despite the fact that the top water level was reached two days later.

Despite this natural disaster, positive emotions are prevailing for the time of the Elbe River Flood. Although words with negative connotations are used, like ‘disaster’, ‘flood’ or ‘crisis’, words with actually positive connotations occur more often, for instance ‘water’ which is nevertheless negative in the present case.

## 5. Potentials and Limits

The emotional data extracted from the metadata of Flickr and Panoramio photos have the potential to enable temporal analyses regarding long-term trends, periodic events and single events. Furthermore patterns of spatial-temporal emotions can be identified as emotional hotspots. Limiting are the existence as well as the popularity of the photo platforms Flickr and Panoramio as they are existing since 2004/2005 and as their popularity is not steady.

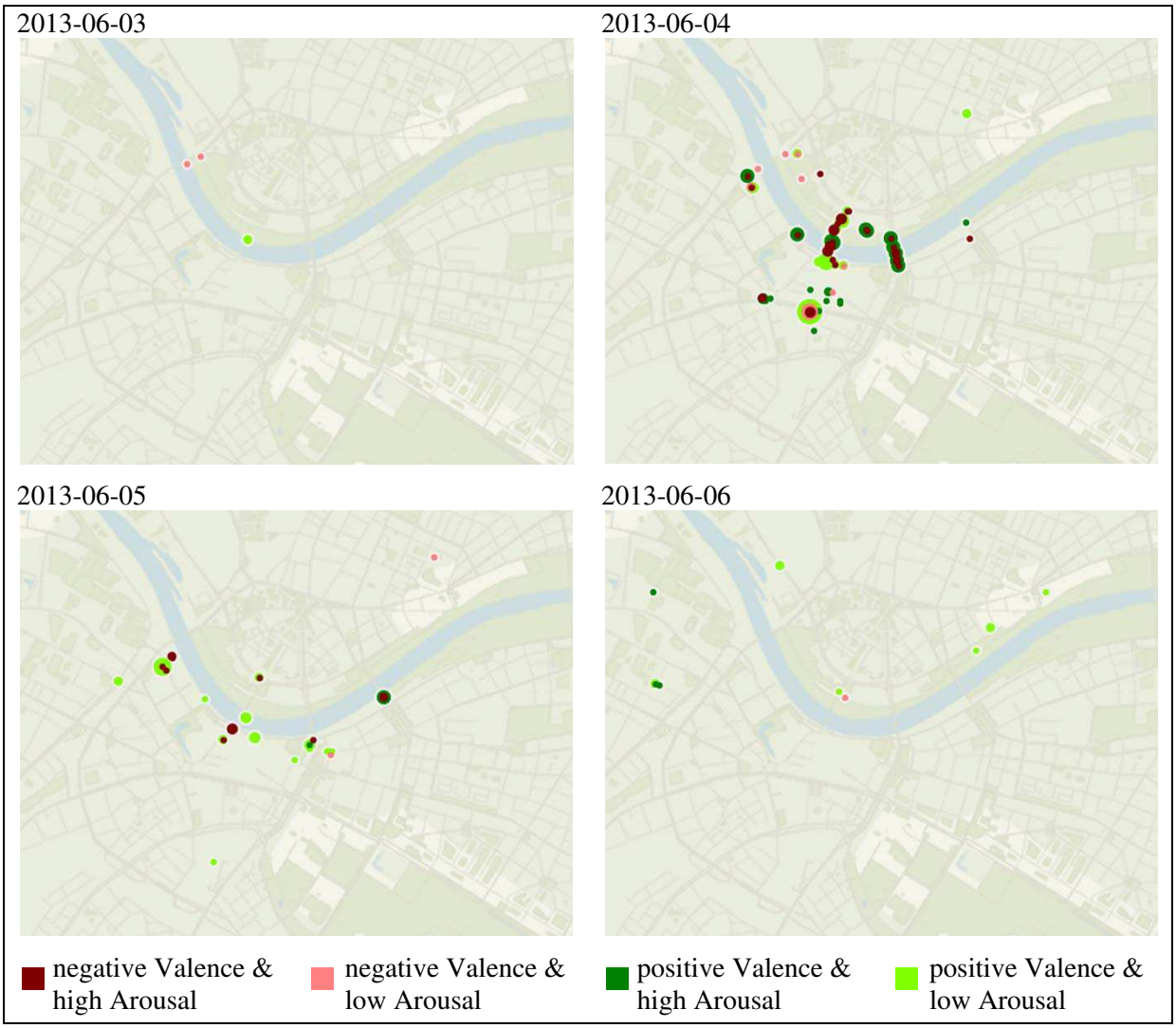



Figure 4. Spatial-temporal distribution of emotions in the first days of the flood in June 2013. Most pictures were taken from the bridges on the 4th June.

The analysis of long-term trends used a time scale of one year but the investigation of time-dependency is not exploited yet regarding the temporal granularity, e.g. the same investigations can be done for weekdays, daytime etc.

The analysis of periodic events works quite well in terms of the number of photos and emotions as an indicator for a periodic event.

The investigation of single events shows that the occurring kind of emotions is not necessarily appropriate since places as well as words can be associated with different emotions, like the word ‘water’. This word is basically positive but can be also negative in a certain context but ANEW and BAWL-R as underlying data do not consider this phenomenon. Thus a general restrictive weak point of the algorithm for extracting emotions from photo metadata is pointed out.

Another, still unaddressed temporal aspect is the problem how to handle a photo with one time stamp but containing emotions of several dates within the text of the metadata.

## References

Bradley M and Lang P, 2010, Affective norms for English Words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-2. The Center for Research in Psychophysiology, University of Florida, Gainesville.

Hauthal E and Burghardt D, 2013, Extraction of location-based Emotions from Photo Platforms. In: Krisp J (ed.), *Progress in Location-Based Services, Lecture Notes in Geoinformation and Cartography XXXV (2013)*,Springer, 3-28.

Hayakawa S, 1952, *Language in Thought and Action*. George Allen & Unwin, London.

Izard C, 1977, *Human Emotions*. Plenum Press, New York.

Russell J, 1980, A circumplex model of affect. *Journal of Personality and Social Psychology*, 39: 1161–1178.

Schimmack U, 1999, Strukturmodelle der Stimmungen: Rückschau, Rundschau und Ausschau. *Psychologische Rundschau*, 50(2): 90–97.

Võ M L-H, Conrad M, Kuchinke L, Hartfeld K, Hofmann M F and Jacobs A M, 2009, The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, 41(2): 534–538.

# Comparing different crowdsourced data: an analysis of Flickr elements, qualities and activities with Geo-Wiki land cover

Alexis Comber[1] and Ross Purves[2]

[1]Department of Geography
University of Leicester, UK
Email: ajc36@le.ac.uk

[2]Department of Geography
University of Zurich, Switzerland
Email: ross.purves@geo.uzh.ch

## 1. Introduction

There is much interest in crowdsourced data and recent applications using such data range from astronomy to zoology (Foody, et al., 2014). The high and increasing volumes of data generation are driven by the ability of citizens to capture geo-referenced information about their daily lives and the environment they experience, using GPS- and web-enabled digital devices (e.g. digital cameras, smartphones, tablets, etc). The result is high volumes of low cost spatially referenced data describing all sorts of phenomenon and processes and consequent interest within the scientific community in using such data.

Much of the initial crowdsourced data research focused on its use to validate other data. The term *'crowd-sourcing'* originally referred to the ability of citizens to validate and correct the errors that an individual might make and to potentially arrive at some *truth* (Goodchild & Li, 2012). A recent example is the Geo-Wiki project. This is a web-based interface to Google Earth that was used to validate a global biofuels availability dataset and (Fritz, et al. 2012) and resulted in a number of other campaigns. Recent work has focussed on the *quality* of the crowdsourced data itself (Comber et al., 2013; See et al., 2013; Foody et al., 2014), especially within the context of using such data to augment or replace data collected under the *designed experiment* (Myers et al, 2010).

However, one of the critical but as yet under-examined aspects of crowdsourced data quality relates to the reasons behind variations in responses. Specifically little work has linked well-known variations in semantics, conceptualisations and cognitive processes with *what* is recorded by the crowd. This is particularly relevant to crowdsourced land cover categorisation activities such as Geo-Wiki as the social construction of spatial data is well-recognised and easily exemplified through land cover (Comber et al., 2005). A long body of research under the banner of *cognitive spatial information theory* describes how the geographic external reality is differently conceptualised and how features are categorized in different ways, under different processes and in different contexts. Classically this included notions of geographic objects and their boundaries (eg Smith and Mark, 2001) and more recently has included ethnophysiography studies, (eg Derungs et al., 2013; Mark and Turk 2003). Much of this early work was concerned with data integration rather than crowdsourced data *per se*. However, its salience is greater now in the context of incorporating crowdsourced data into scientific analyses. This in turn suggests a considerable gap in current understanding: how to scale up the results from the often small scale, but tightly controlled studies (for example examining the spatial concepts used by relatively small numbers of people in specific groups) to mass observation activities such as Geo-Wiki.

This research explores how landscape features captured by the Geo-Wiki Human Impact campaign (Comber et al., 2013) relate to other crowdsourced data describing landscape available from Flickr (Purves et al., 2011; Hollenstein and Purves, 2010). It seeks to quantify the relationships between the semantics captured by Geo-Wiki and Flickr and to identify

uncertainties associated with divergent semantics. In so doing it seeks to identify methods with which to scale up previous work in cognitive spatial information theory.

## 2. Methodology

The idea was to integrate data from the Human Impact Geo-Wiki with concepts derived from Flickr images. Then to examine how the landscape concepts captured by Geo-Wiki land cover labelling, relate to and vary against terms used as tags to describe georeferenced Flickr images.

Since many tags are typically toponyms or other specific information (Sigurbjornsson et al. 2008), the taxonomy of tags developed by Purves et al (2011) was used. This seeks derive place-related facets from images based on two georeferenced collections in the UK. The facets are termed *elements, qualities* and *activities* and have been shown to be useful in characterising locations (Edwardes and Purves, 2007). In total some 581 terms are in the taxonomy, including 313 elements, 107 qualities and 161 activities.

The Human Impact Geo-Wiki generated ~65,000 data points describing a land cover at randomly selected locations. A number of scientific papers fully describe this data set (Comber, et al., 2013; See et al., 2013 ), but in brief, volunteers were asked to allocate an area covering 30 arc seconds to one of 10 land cover classes using the Google Earth interface.

The Geo-Wiki captured 293 data points in the UK and Ireland and 5239 in the USA. For each of these a 30 arc second search window was used to extract the coincident Flickr data. Searching through 28 million Flickr images associated with the UK alosn, 90 Geo-Wiki points in the UK and Ireland and 91 in the USA were identified with one or more Flickr images. We then listed all of the elements, qualities and activities, as well as the total number of images and unique Flickr users, associated with this Geo-Wiki point.

A *Latent Dirichlet Allocation* (LDA) ( Blei et al., 2003) was used to analyse the content of the tags associated with each class, in each region. LDA seeks to explain similarity using unobserved, latent groups or *topics*. The idea is that each *bag of words* includes a number of embedded topics which are indicated here by the tags associated with land-cover classes. Latent approaches consider the data (documents) and the hidden concepts they contain (topics) from the standpoint of naivety and seek to determine the underlying similarities between terms and concepts. Here, the Flickr elements, qualities and activities, linked to the Geo-Wiki land cover data were used to create a document corpus, with a document for each land cover class.

## 3. Results and Interpretation

LDA analysis was run on the corpus using the *topicmodel* package [2] in R. Six latent variables or topics were specified, and are characterised by the terms that are most strongly associated with them from the posterior probabilities generated by the LDA of each term being associated with each topic (Figure 1). For example, in the UK this suggests that there are 3 distinct topic groups: Topics 1 (*landscape, sky, sunset road, countryside, flowers*), Topics 3 and 5 (*green, village, black clouds, bus, stone, garden, field*, etc) and Topics 2, 4 and (*white, nature, landscape, snow, dog, sky, church, sunset,* etc). It is evident that in general the posterior probabilities are much lower in the US data compared to the UK data.

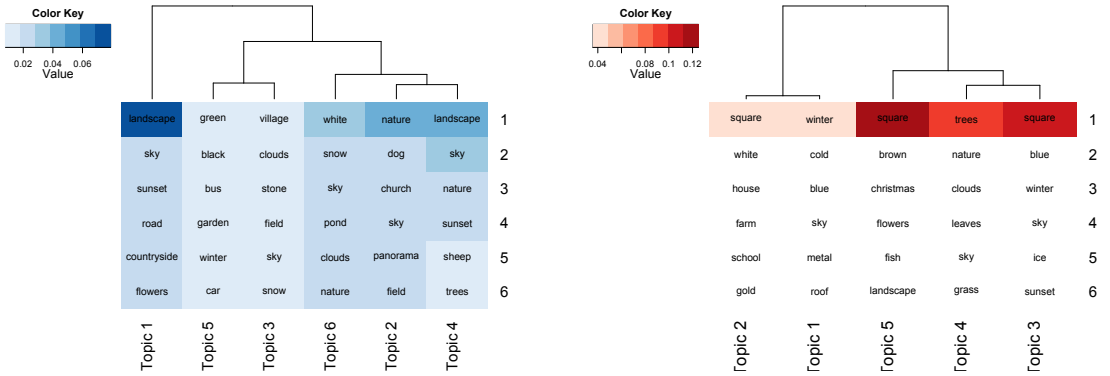



Figure 1. The terms most strongly associated with each topic in the UK (blue) and in the USA (red), shaded by the posterior probability of belonging to that topic and with the topics clustered.

The LDA generates posterior probabilities that each land cover class (document) is associated with each topic via their semantics. The relationships between topics and land cover class can be visualised in a network, where the edges are defined by probability. Edges (connections) between classes and topics (vertices) were removed if the posterior probability for each topic-year pair was less than 0.1 and remaining connections between topics and land cover classes are shown in Figure 2. This shows that in the UK Topics 1, 2 and 3 are uniquely associated with Class 4, *Cultivated,* Topic 4 with Classes 3, 5 and 9 (*Grassland, Mosaic* and *Barren*), Topic 5 with Classes 1 and 7 (*Trees* and *Urban*) and Topic 6 with Classes 3, 7 and 9 (*Grassland, Urban* and *Barren*). In the USA there is much more disconnection between the classes, which may be explained by the low tag volume. It is also possible to explore the connectedness of the semantics associated with different land cover classes. Further examination of Figure 2 is possible. For example, in the UK Class 4, *Cultivated*, is not semantically connected with the other classes, that Classes 1, 7 and 9 (*Trees*, *Urban* and *Barren*) are similar, that Class 5 (*Mosaic*) is weakly connected to Class 9 (*Barren*) and that class 3 (*Grassland*) is strongly connected to Classes 5 and 7 (*Mosaic* and *Urban*).

## 4 Discussion Points

A number of areas for future consideration have been identified in this initial exploratory work. First, though the amount and variety of crowdsourced data has rapidly increased in recent years, we could still only identify Flickr images related to ~1/3 of Geo-Wiki points (from an initial set of 28 million). Second, there are research gaps in how to integrate such data, particularly with respect to a quantitative handle on the semantics of the crowd. For example, in tag lists, there can be problems in resolving ambiguity as the context in which an individual tag to use. For example, some tags may be both elements and qualities (e.g. city is both an element in and of itself, and a quality with respect to buildings). Thirdly, recent research is clearly drawing from a much wider range of data sources, labelled in different and novel ways, potentially reflecting the rapid increase in the platforms and systems available to individual citizens that enable them capture and share a diverse range of different types of information, describing the world we live in. There are obvious areas for future research in considering who contributes such data, the impact of digital divides on the nature of the information that is contributed and potential biases towards western, developed populations and of course the nature of the technologies used to capture and share such information. On-going work is considering these issues.

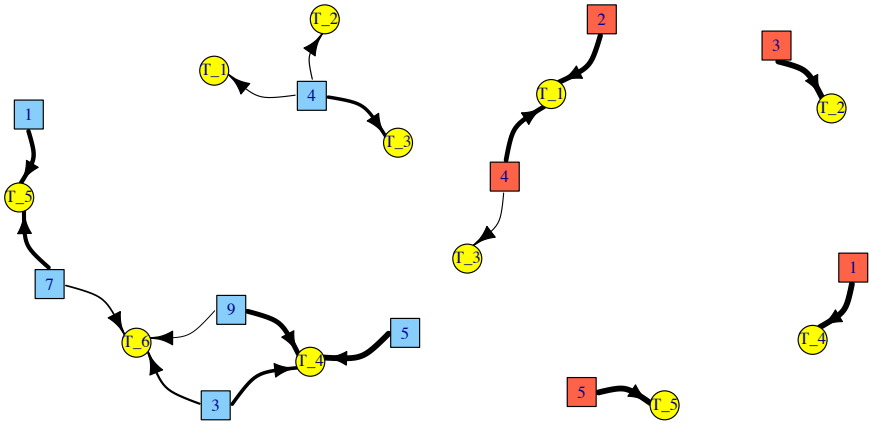



Figure 2. The links between topics and land cover classes, with the strength of the link as defined by the posterior probability as determined by the LDA model indicated by the edge widths. The Uk data are in Blue and the USA in Red. (NB Class 1 *Trees*, Class 3 *Grassland*, Class 4 *Cultivated*, Class 5 *Mosaic*, Class 7 *Urban*, Class 9 *Barren*).

## References


Blei DM, Ng AY and Jordan MI, 2003, Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993-1022.

Comber AJ, Fisher PF and Wadsworth RA, 2005, What is land cover? *Environment and Planning B: Planning and Design*, 32:199-209.

Comber A, See L, Fritz S, Van der Velde M, Perger C and Foody GM, 2013, Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation and Geoinformation*, 23:37–48.

Derungs, C, Wartmann, F, Purves, R S, and Mark, D M (2013) The meanings of the generic parts of toponyms: use and limitations of gazetteers in studies of landscape terms. In *Spatial Information Theory* (pp. 261-278).

Edwardes AJ and Purves RS, 200, A theoretical grounding for semantic descriptions of place. In *Web and Wireless Geographical Information Systems* (pp. 106-120). Springer Berlin Heidelberg.

Foody GM, See L, Fritz S, Van der Velde M, Perger C, Schill C, Boyd DS and Comber A, 2014, Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality. *The Cartographic Journal* DOI: http://dx.doi.org/10.1179/1743277413Y.0000000070

Fritz S, McCallum I, Schill C, Perger C, See L, et al (2012) Geo-Wiki: An online platform for improving global land cover. Environmental Modelling & Software 31: 110–123. doi:10.1016/j.envsoft.2011.11.015

Goodchild MF and Li L, 2012, Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110-120.

Hollenstein L, and Purves R 2014, Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, (1), 21-48.

Jockers ML, 2013, *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Mark, DM and Turk, AG, 2003, Landscape categories in yindjibarndi: Ontology, environment, and language. In: Kuhn, W., Worboys, M.F., Timpf, S. (eds.) COSIT 2003. LNCS, vol. 2825,pp. 28–45. Springer, Heidelberg

Myers JL, Well A, Lorch RF, 2010, *Research design and statistical analysis*, Routledge, New York

Purves, R, Edwardes, A, and Wood, J, 2011, Describing place through user generated content. *First Monday*, *16*(9).

See, L, Comber, AJ, Salk, C, Fritz, S, Van der Velde, M, Perger, C, Schill, C, McCallum, I, Kraxner, F and Obersteiner M, 2013, Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. PLoS ONE, 8(7): e69958. doi:10.1371/journal.pone.0069958.

Sigurbjornsson B and Van Zwol R, Flickr tag recommendation based on collective knowledge. In WWW '08: Proc. 17th international conference on World Wide Web (Beijing, China, 2008), ACM, pp. 327–336..

Smith B and Mark D, 2001,Geographical categories: an ontological investigation. *International Journal of Geographical Information Science,* 15:591-612.

# Towards a framework for automatic geographic feature extraction from Twitter

Enrico Steiger, Johannes Lauer, Timothy Ellersiek, Alexander Zipf

GIScience Research Group, Institute of Geography, University of Heidelberg, Berliner Straße 48 D-69120 Heidelberg
Email: {enrico.steiger; johannes.lauer, timothy.ellersiek,zipf}@geog.uni-heidelberg.de

## 1. Introduction

Interactive social media platforms offer a tremendous amount of volunteered, user-generated content (Flickr, Twitter, etc.). Together with volunteered geographic information (VGI) they potentially provide a valuable source of information which is increasingly recognized, but particularly in GIScience not utilized to its full potential. Twitter as one location-based social network in particular, provides the ability to sense geo-processes and to gain knowledge about the individual user perception towards geographic objects. A georeferenced tweet represents a proxy of a real world observation and contains spatial, temporal and semantic information. These social sensor measurements depend on particular tweet locations and are influenced by the individual user perception of urban space. Although there is a growing research body conducting Twitter analysis, a key challenge remains whether this noisy biased data source forms a representative sample for the knowledge discovery of geographic information. Location information retrieved from Twitter data is spatio-temporally and semantically uncertain. One of the main research aims is therefore to investigate whether geographic features from tweets can be detected and extracted. Furthermore, we explore whether the inferred geometries of features match with real world spatial objects (e.g. points of interest).

In this work we propose a framework to infer geographic features from unstructured georeferenced Twitter data using semantic topic modelling and spatial clustering techniques. Given the detected and extracted geographic features from Twitter, we applied a geometry computation and compared the results with map features from OpenStreetMap.

### 1.1 Related Work

There are a number of previous studies on a macroscopic scale aiming to infer direct or indirect geographic information from Twitter using provided metadata, the semantic tweet content or geographic coordinates. Cha et al. (2010) focus on enriching georeferenced tweets by inferring the location from user profiles and in addition their social network. Gonzalez and Chen (2012), Hiruta et al. (2012) and Lee and Hwang (2012) further develop a location inference system using user profile location, semantic classified tweet content or GPS coordinates from the geotag. Hong et al. (2012) develop a location aware topic model to correlate relationships between location and words. Dalvi et al. (2012) geolocate users by matching posted tweets containing indirect spatial information to real world spatial objects. Sengstock and Gertz (2012) introduce a framework for unsupervised extraction of latent geographic features from georeferenced Flickr data.

## 2. Methods

Tweets represent a spatio-temporal signal with a semantic information layer. We have extracted a semantic dimension over geographic space in order to infer geographical features on a small map scale (street level).

## 2.1 Dataset

For our case study we use a dataset only containing geotagged tweets from the area of Greater London. Table 1 shows some further details regarding the retrieved Twitter data.

| Dataset | Greater London (UK) |
|---|---|
| Bounding Box (WGS 84) | -0.5543,51.2386,0.3038,51.731 |
| Timespan | 01/10/2013-31/03/2014 |
| Covered Area | 3265387 km² |
| Number of geotagged tweets | 15.8 million |
| Number of tweeted User | 433555 |

Table 1: Meta information for our selected Twitter dataset

## 2.2 Framework

All tweets are collected in real-time through the official Twitter streaming API (https://dev.twitter.com/docs/api/streaming). The semantic tweet content from every user is then preprocessed to remove whitespaces, punctuations and numbers. In the next step all tweet corpora from Twitter undergo a natural language processing step by applying tokenization, stemming and stop word filtering (Lewis et al. 2004). We are using latent dirichlet allocation (LDA) as one semantic probability based topic extraction model introduced by Blei et al. (2003). The unsupervised machine learning model identifies latent topics and corresponding word clusters from our large collection of tweets. This technique reduces the semantic dimensions and works efficiently especially on large unseen datasets. It is a sophisticated method compared to arbitrary simple keyword filtering techniques which have limited scalability. Figure 2 shows an exemplary LDA probabilistic topic extraction visualization for the highest assignments (>0.3) for the topic associated words “trafalgar” and “square”. The words “photo”, “london” and “england” also appear and show lower topic assignments (<0.3). As a result, high density areas of topic relevant classified tweets are closer to the real world object Trafalgar Square.


Figure 1: LDA topic association indicator for words “trafalgar” and “square” over all topic related filtered georeferenced tweets in London (n = 3796).

After the tweets have been processed and classified with LDA topic modelling, we chose DBSCAN (Ester et al. 1996) as a density based point clustering and classification algorithm to process the point cloud data. The algorithm detects dense clusters and filters noisy points. From the densest cluster where most tweets have been assigned to, we generate a trajectory which can be compared and matched with the corresponding geographic object from OpenStreetMap (OSM).

# 3. Results

## 3.1 Point Clustering

DBSCAN is applied in order to detect statistically significant semantic and geographic centroids of LDA classified tweets for the topic "oxford street". The Euclidean distances between the topic associated tweets have been normalized. The minimum number of points to form a cluster was defined to be 7 with a density reachability distance of ε=0.1. As a result (Figure 2) all tweets in cluster 1 (91% of total extracted features) are density connected points without scattering. This cluster is spatially concentrated along the real world geographical object Oxford Street. Associated tweets (n=1085) are our targeted point cluster. Cluster 2 and 3 are locally occurring dispersed clusters, showing a low density-reachable tweet distribution with a low amount of associated features.

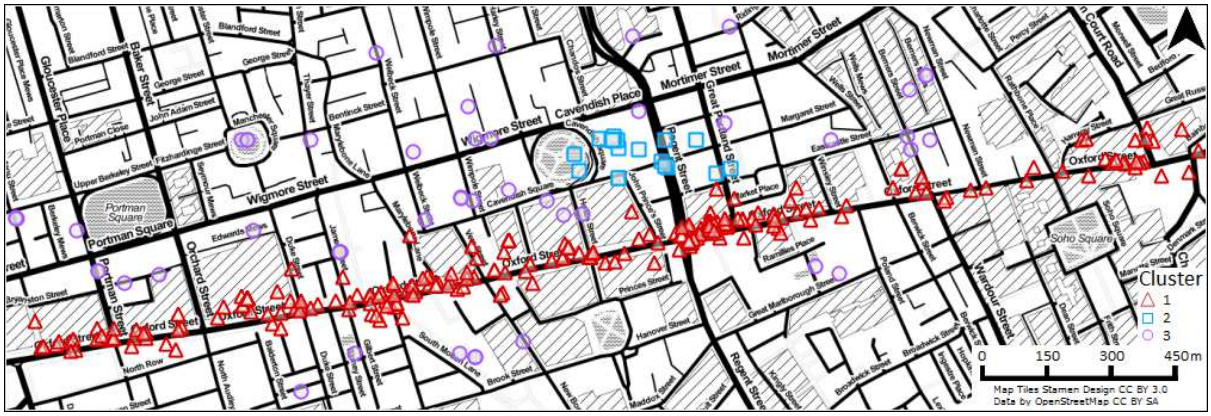



Figure 2: LDA topic associated words "oxford" and "street" for georeferenced tweets in London after applying DBSCAN clustering (n = 1186).

## 3.2 Geometry Extraction and OSM Feature Comparison

In order to extract geometric features and compare them with an existing map, several processing steps have to be taken. The first step is the extraction of the corresponding geometry for the new feature. In our case we created a linestring by applying the principal curve algorithm of Hastie and Stuetzle (1989), which is able to fit a line string to an unsorted point data set. The result is a geometric representation of Oxford Street (Figure 3). The second step is to match the new generated linestring with the corresponding feature from the OSM road network. As a quality indicator for the positional accuracy, the Hausdorff distance is calculated. For both linestrings, the Hausdorff distance is 0.0030468 which provides an indication for their similarity.

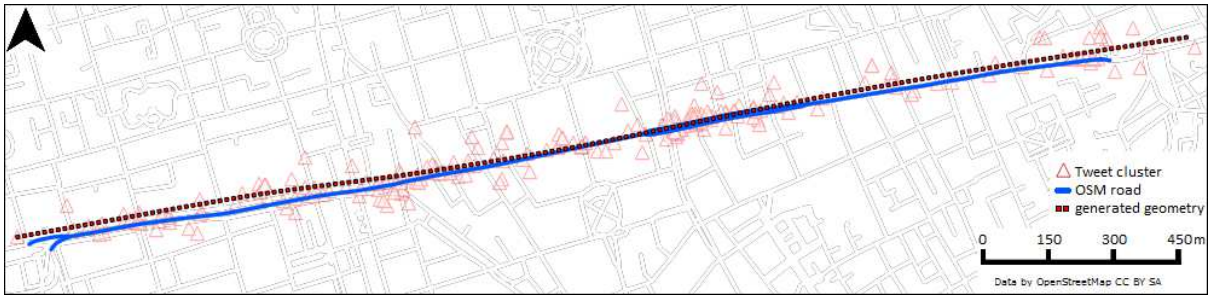



Figure 3: Oxford Street - extracted linestring geometry from tweets and comparison with OSM road (Hausdorff distance = 0.0030468)

## 4. Conclusion

Our results for the selected case study in London show that geographic features can be successfully extracted from Twitter by using geographic and semantic information. We were able to generate a new road feature from Twitter measurements which is quite similar to the mapped OpenStreetMap feature. Limitations of the study include the geographic objects themselves which might be too complex to be clearly detected from the spatial-semantic signal, or the tweet signal might not be significant enough and too sparse to be detected at all.

## References

Blei, D., Ng, A. and Jordan, M., 2003, Latent dirichlet allocation. Journal of machine Learning research, 993–1022.

Cha, Meeyoung, et al., 2010, Measuring User Influence in Twitter: The Million Follower Fallacy. In: ICWSM 10, 10-17.

Dalvi, N., Kumar, R. and Pang, B., 2012, Object matching in tweets with spatial models. In Proceedings of the fifth ACM international conference on Web search and data mining - *WSDM '12*. New York, USA, 43.

Ester, M. et al., 1996, A density-based algorithm for discovering clusters in large spatial databases with noise. KDD, 96.

Gonzalez, R. and Chen, Y., 2012, TweoLocator: A Non-Intrusive Geographical Locator System for Twitter. In: Proceedings of the 5th International Workshop on Location-Based Social Networks, 24–31.

Hastie, T. and Stuetzle, W., 1989. Principal Curves. Journal of the American Statistical Association, 84(406), 502–516.

Hiruta, S. et al., 2012, Detection , Classification and Visualization of Place-triggerd Geotagged Tweets. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing.

Hong, L. et al., 2012, Discovering geographical topics in the twitter stream. In: Proceedings of the 21st international conference on World Wide Web - *WWW '12*. New York, 769.

Kinsella, S., Murdock, V. and Hare, N.O., 2011," I ' m Eating a Sandwich in Glasgow ": Modeling Locations with Tweets. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents, 61–68.

Lee, B. and Hwang, B.-Y., 2012, A Study of the Correlation between the Spatial Attributes on Twitter. 2012 IEEE 28th International Conference on Data Engineering Workshops, 337–340.

Lewis, D.D. et al., 2004, RCV1: A New Benchmark Collection for Text Categorization Research. The Journal of Machine Learning Research, 5, 361–397.

Sengstock, C. and Gertz, M., 2012, Latent geographic feature extraction from social media. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems - *SIGSPATIAL '12*. New York, USA, 149.

# Mapping spatial uncertainty in object-fields: the case of site suitability analysis

Thomas J. Cova[1], Piotr Jankowski[2]

[1]University of Utah, Department of Geography
260 South Central Campus Dr. Rm 270
Salt Lake City, UT 84112-9155
Email: cova@geog.utah.edu

[2]San Diego State University, Department of Geography
San Diego, CA 92182-4493
Email: pjankows@mail.sdsu.edu

## 1. Introduction

Land suitability analysis begins with a set of spatial factors and uses a combination rule to assign each land unit a suitability score (Malczewski 2006). While this is very valuable in identifying suitable regions, it holds short of delimiting an explicit site boundary. Site suitability analysis, also referred to as land acquisition (Wright et al. 1983) or site allocation modeling (Brookes 1997), begins with objectives regarding desired site characteristics and returns the set of land units that comprise an ideal site. Objectives may include maximizing suitability or minimizing cost, and geometric constraints are often placed on site area, shape, and contiguity. A common goal is generating good candidates for a decision maker rather than a single best site.

Incorporating uncertainty into the decision-making process is a challenge in site suitability analysis (Aerts et al. 2003). Uncertainty can enter through the quality of the input data, the decision weights, or the models and parameters that make up the definition of suitability. Recently, new methods were developed for mapping the spatial variation in land suitability uncertainty (Ligmann-Zielinska and Jankowski, 2014). A suitability map analyzed in concert with an uncertainty map can be used to find areas that are high in suitability and low in uncertainty. This advancement may also help forward uncertainty representation in site suitability analysis, although uncertainty must be conceptualized in a different manner given the combinatorial number of possible sites.

The goal of this research is to improve the representation of uncertainty in site suitability analysis. We combine new methods for modeling uncertainty in land suitability analysis with object-fields to incorporate uncertainty into optimal site modeling. Although uncertainty can enter the process at many points, our focus is the factor-weights provided by decision makers. The primary research question is: how does uncertainty in the suitability factor-weights affect uncertainty and sensitivity in the resulting candidate sites?

## 2. Methodology

One approach to representing uncertainty in site suitability modeling is to rely on a field of spatial objects (Cova and Goodchild, 2002). An object-field assigns each location one or more spatial objects. In a continuous field this leads to an infinite number of locations and associated objects, but in the context of a finite tessellation it can be reduced to assigning a best site object to each spatial unit. Site uncertainty can be derived by combining the uncertainty values of a site's land units, and sensitivity can be viewed as the stability in a site's characteristics (e.g. suitability, cost, area, shape) under changes in the input data, weights, or parameters.

The problem of finding an ideal site is formulated as a spatial optimization model. The objective is to maximize the sum of the suitability scores for land units that comprise a site. Secondary objectives, such as minimizing cost or maximizing compactness, can be formulated as constraints or incorporated using a weighted objective function. Our search is for a contiguous site with constraints on area and compactness, which is defined using a normalized area-to-perimeter measure *C* that ranges from 0 (sinuous) to 1 (compact):

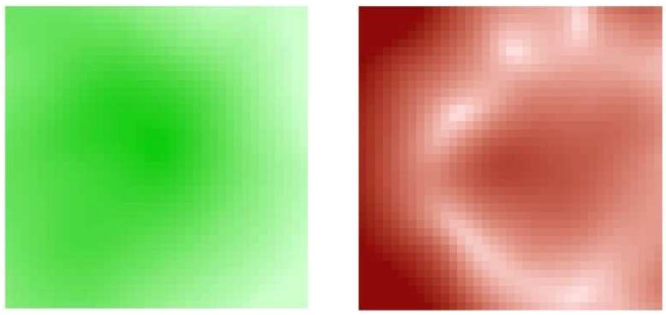
$$\mathrm{C} = \sqrt{A/(.282P)^2} \quad (1)$$

Solving this model for the optimal solution comes with computational overhead, and a region-growing heuristic algorithm is much more efficient for large problems. The heuristic begins at each cell and seeks to maximize site suitability using a greedy approach. At each step, the cell on the site 'fringe' (i.e. cells that share an edge with the current site) is added that most increases the objective, subject to any constraints. To avoid local optima and improve solution quality, we implemented a semi-greedy approach that runs the heuristic *n* times from each cell and randomly selects from the most promising fringe cells (i.e. those within a set percent of the greedy option). This is performed for every cell to yield a field of best sites that possess unique attributes and geometric properties.

A suitability map is derived by combining input map layers using weights that represent the relative importance of suitability factors. Weighted summation, ideal point, or multi-attribute utility functions can be used to calculate a suitability value for each map unit (e.g. raster cell), and we used ideal point. To account for input uncertainty in the factor weights (i.e. decision maker uncertainty about relative importance), an unbiased sample of weights is prepared using a sampling scheme proposed by Sobol (1993) and a probability density function appropriate for the set of factor weights under investigation. A large number of realizations (> 10000) of a suitability map is computed using Monte Carlo simulation, and a mean suitability value for each spatial unit is derived, leading to a mean suitability map and an accompanying standard deviation map called an *uncertainty map*.

## 3. Results

The findings in this section are based on the input mean suitability and uncertainty maps shown in Figure 1 (37 rows by 37 columns). The standardized suitability values range from 0.091 to 0.70 and the uncertainty values range from 0.004 to 0.21 on a 0-to-1 scale.



Mean suitability (0.0-1.0) Uncertainty (0.0-1.0)

Figure 1. Mean suitability (green, more suitable) and uncertainty (red, more uncertain) maps.

The goal of the first experiment was to reveal the trade-off between identifying high-suitability sites and those with less uncertainty. Figure 2 shows that site suitability declines more rapidly as site size increases when uncertainty is more constrained. High uncertainty sites are limited to a mean of 0.10/cell and low uncertainty sites are limited to a mean of 0.032/cell. Thus, the larger the desired site, the more that suitability must be sacrificed to reduce uncertainty, as the heuristic cannot find a large, highly-suitable site with low mean-uncertainty.

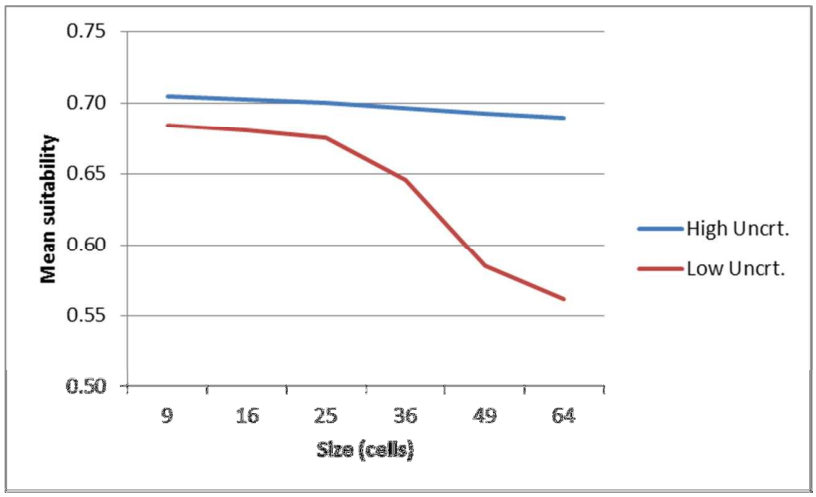
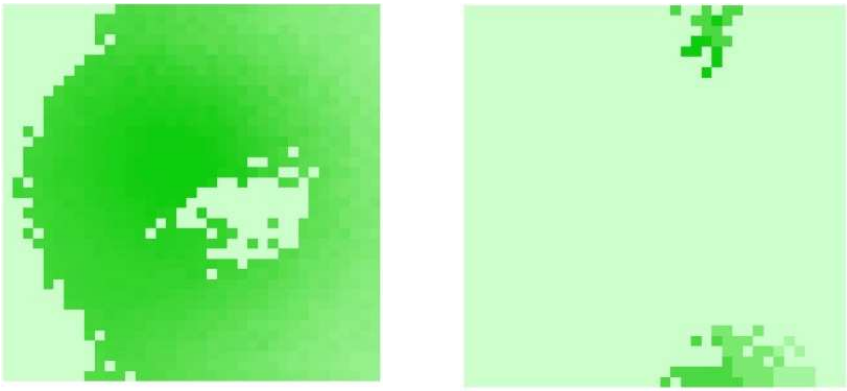



Figure 2. Tradeoff between site suitability and site uncertainty across site size.

In the next experiment, we mapped the variation in the objective value for a 64-cell site. As each cell is a separate problem, the objective is to find the most suitable site that includes the seed cell, subject to constraints on uncertainty (as above) and shape (compactness >= 0.5 using Equation 1). Figure 3 (right) shows that few cells met the low uncertainty tolerance case (green cells), but for the high uncertainty tolerance case (left), there were many cells that resulted in a feasible site, with the best ones shown in darker green.



High uncertainty tolerance (Objective value range: 0-44.1)

Low uncertainty tolerance (Objective value range: 0-28.7)

Figure 3. Mapping the objective value for all cells under high and low uncertainty.

Next, we mapped the spatial sensitivity in the best site for each cell after reducing the uncertainty. For each cell, the high and low uncertainty sites were compared for the percentage

of shared cells. If the reduction in uncertainty resulted in a significant change in the site's footprint, then the spatial sensitivity to uncertainty was considered high. Figure 4 shows the results for a 64-cell site with compactness greater than 0.5, where the attribute mapped is the percentage of cells shared between the high and low uncertainty solutions. The darkest blue cells had solutions that approached 100 percent in shared cells between the high and low uncertainty cases, and the lighter blue cells are example sites with higher sensitivity, as the footprint of the two solutions deviated more.

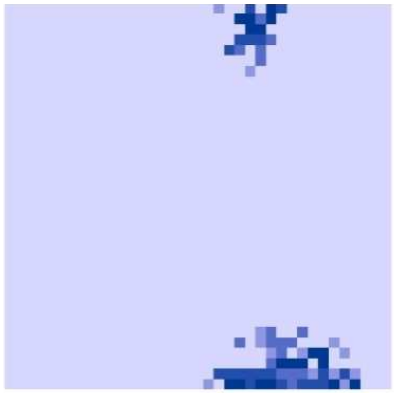


Figure 4. Spatial sensitivity in the site footprint at each cell with uncertainty reduction (0-100%).

## 4. Conclusion

A new method for incorporating factor-weight uncertainty into site suitability analysis was presented. The results show that a decision maker seeking a highly-certain solution may be faced with far fewer options that meet the site criteria than the case of relaxing the uncertainty threshold. Second, the identification of good sites depends to a large degree on the initial spatial correlation in the suitability and uncertainty input maps. If highly-suitable, relatively-certain areas are present in these maps, then a site search algorithm will identify good candidates, even for decision makers with a lower tolerance for uncertainty.

## References

Aerts, J.C.J.H, Goodchild, M.F., and Heuvelink, G.B.M. (2003) Accounting for spatial uncertainty in optimization with spatial decision support systems. *Transactions in GIS*, 7(2): 211-230.

Brookes, C.J. (1997) A parameterized region-growing programme for site allocation on raster suitability maps. *International Journal of Geographical Information Science*, 11(4): 375-396.

Cova, T.J., and Goodchild, M.F. (2002) Extending geographic representation to include fields of spatial objects. *International Journal of Geographical Information Science*, 16(6): 509-532.

Ligmann-Zielinska, A., and Jankowski, P. (2014) Spatially-explicit integrated uncertainty and sensitivity analysis of criteria weights in multicriteria land suitability evaluation. *Environmental Modeling & Software, 57: 235-237.*

Malczewski J, 2006, GIS-based multicriteria decision analysis: A survey of the literature. *International Journal of Geographical Information Science*, 20: 703-726.

Sobol, I. M. 1993. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment,* 1, 407-414.

Wright, J., Revelle, J., and Cohon, J. (1983) A multi-objective integer programming model for the land acquisition problem. *Regional Science and Economics*. 13(1): 31-53.

# On Use of Fuzzy Surfaces to Detect Possible Elevation Change

Peter Fisher[1] and Jan Caha[2]

[1]Department of Geography, University of Leicester, Leicester, LE1 7RH, UK
Email: pff1@leicester.ac.uk,

[2]Department of Geoinformatics, Palacký University, 17 Listopadu 50,771 46 Olomouc, Czech Republic
Email: jan.caha@klikni.cz

## 1. Introduction

A fuzzy surface (or Digital Elevation Model, DEM) is a model that captures both the surface elevation itself and its uncertainty. In a classic (crisp) surface for each coordinate $x, y$, a value, $z$, is stored that represents the "height" of the surface. On the other hand for each coordinate a fuzzy surface stores a fuzzy number, $\tilde{Z}$, that represents the possible range of values that the surface can have at this location (Lodwick et al. 2008). A fuzzy number is a special case of a fuzzy set that represents vague, imprecise or ill-know values (Dubois and Prade 1983). Fuzzy numbers are often compared to probability distributions, but although there are some similarities they are not interchangeable. Probability originates in a stochastic process of variability while fuzziness is the result of imprecision, vagueness or lack of knowledge. Lodwick et al. (2008) argue that actually most of the uncertainty associated with geographical surfaces is fuzzy a similar premise to that of Loquin and Dubois (2010).

In this paper we examine elevation change over a British coastal dune field which is essential for coastal flood defence providing environmental security. We aim to detect elevation change between fuzzy surfaces constructed from multi-temporal LiDAR data. Such change is calculated by means of fuzzy arithmetic and the resulting differential surface is classified into categories by the amount of change with use of possibility theory. The main focus is on methods, the case study being used as an illustrative example.

## 2. Fuzzy Surface

In a standard raster Digital Elevation Model each grid cell records a single number which is usually intended to record the elevation at the center of the cell, but could be representative of elevations within the cell in some other way. In a fuzzy surface the single or Boolean value is replaced by a fuzzy number (Loquin and Dubois 2010). The simplest form of a fuzzy number is as a Triangular Fuzzy Number (TFN). This indicates the support or spread of the number and the core or peak of membership. In a TFN the core is infinitesimally small where membership or degree of the number in the set of numbers equal to the target number is exactly 1.

The fuzzy surfaces (Fig. 1) for the current study were inferred from LiDAR data with a 2m grid size. The same DEM was previously studied by Fisher et al. (2007). Here we generate the fuzzy surfaces by applying simple filters; specifically minimum, median and maximum filters over a 5x5 moving window, in order to obtain values to describe the TFN.

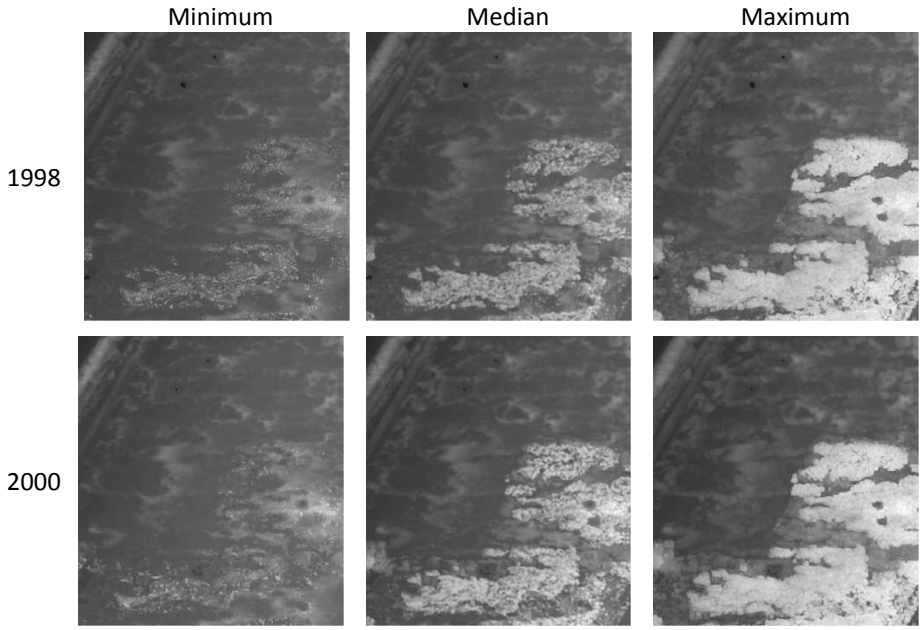



Figure 1: Minimum, median and maximum values describing TFNs for every grid cell in the DEM for the dunefield in 1998 and 2000

## 2. Change detection

Having represented the elevation as two fuzzy surfaces, each pixel of these surfaces can be indicated by the triplets [a,b,c] (surface from year 2000) and [d,e,f] (surface from year 1998), then the difference between the two is another fuzzy number (Lodwick et al., 2008):

$$[a, b, c] - [d, e, f] = [a - f, b - e, c - d]. \quad (1)$$

For the location illustrated in Fig. 2, the crisp difference is considerable, 10.51m, but the fuzzy difference is [-14.12, 0.61. 14.82]m (Fig. 3); the difference in the median heights is very small although the range of possible values is quite large and includes the value of the the crisp result.

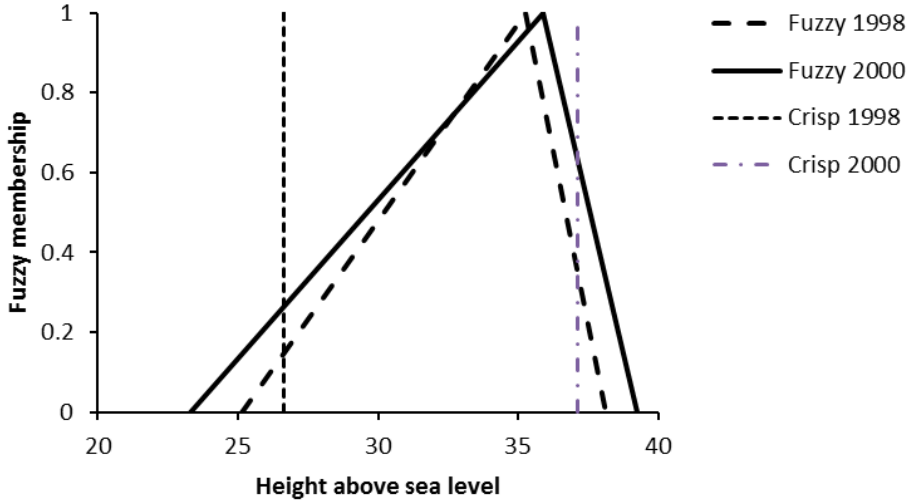



Figure 2: Point values of elevations represented as crisp values and TFNs for 1998 and 2000

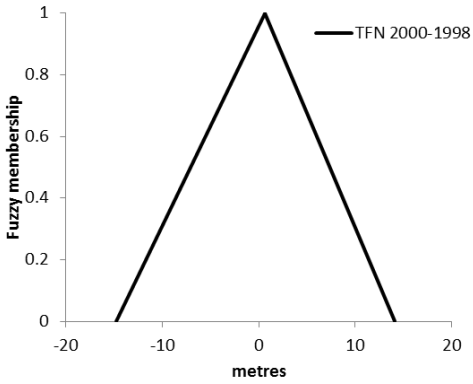



Figure 3: TFN representing the change in elevations between 2000 and 1998.

The resulting comparison of the fuzzy surfaces represented as the TFN is shown in Fig. 4.

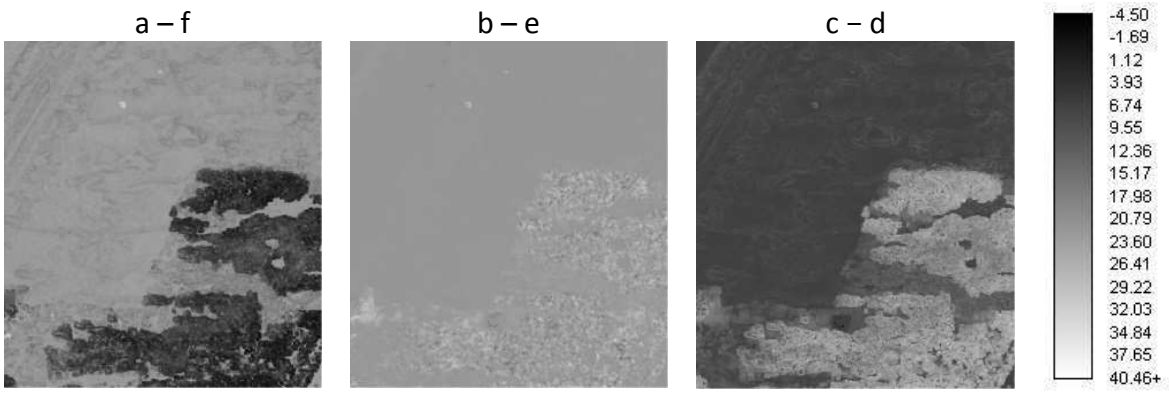



Figure 4: The TFN representing the difference in elevations between 1998 and 2000

## 3. Evaluation of the Fuzzy surface

A simple operation that can be performed on any pair of surfaces is to identify those areas where the surface is greater than and less than a specified threshold $t$. Compared with working with a crisp surface, this process is more complex working with a fuzzy surface. In some cases it cannot be said if the fuzzy number is strictly smaller or greater than the threshold. For example, to determine whether the dune field has changed elevation we can compare the TFN in Figure 3 with the threshold $t = 0$. In this case the threshold is actually contained within the fuzzy number illustrated in Figure 3, and so it is hard to specify if the fuzzy number is greater than or less than 0.

A system that allows comparison of fuzzy numbers and crisp values is proposed by Dubois and Prade (1983). This system utilizes Possibility theory (Dubois and Prade 1986) and provides two measures, possibility ($\Pi$) and necessity ($N$). By definition the possibility of an event is always greater than or equal to its necessity, because necessity is the stricter measure (Dubois and Prade 1986). Complete mathematical foundations of Possibility theory can be found in Dubois and Prade (1986). In GIS Possibility theory has been used for modelling soft queries on crisp data (Caha et al. 2014).

Suppose that we have a TFN [a,b,c] defined by a membership function:

$$\mu_{\tilde{A}}(x) = \begin{cases} 0, \ if \ x < a \\ \frac{x-a}{b-a}, \ if \ a \leq x \leq b \\ \frac{c-x}{c-b}, \ if \ b \leq x \leq c \\ 0, \ if \ c < x \end{cases} \quad (1)$$

In the case of a fuzzy surface where every cell $\tilde{C}_{i,j}$ of the grid is a fuzzy number, then to compare these fuzzy numbers to the threshold $t$ we will obtain the possibility and necessity of $\tilde{C}_{i,j} < t$ (Eqs. (2 and 3)) and $\tilde{C}_{i,j} > t$ (Eqs. (4 and 5) and the result is shown in Fig. 5.

$$\Pi_{\tilde{C}_{i,j}<(t)} = \begin{cases} 0, & if\ t < a \\ \frac{t-a}{b-a}, & if\ a \le t \le b \\ 1, & if\ b < t \end{cases} \quad (2)$$

$$N_{\tilde{C}_{i,j}<(t)} = \begin{cases} 0, & if\ t < b \\ 1 - \frac{c-t}{c-b}, & if\ b \le t \le c \\ 1, & if\ c < t \end{cases} \quad (3)$$

$$\Pi_{\tilde{C}_{i,j}>(t)} = \begin{cases} 1, & if\ t < b \\ \frac{c-t}{c-b}, & if\ b \le t \le c \\ 0, & if\ c < t \end{cases} \quad (4)$$

$$N_{\tilde{C}_{i,j}>(t)} = \begin{cases} 0, & if\ t < b \\ \frac{c-t}{c-b}, & if\ a \le t \le b \\ 0, & if\ b < t \end{cases} \quad (5)$$

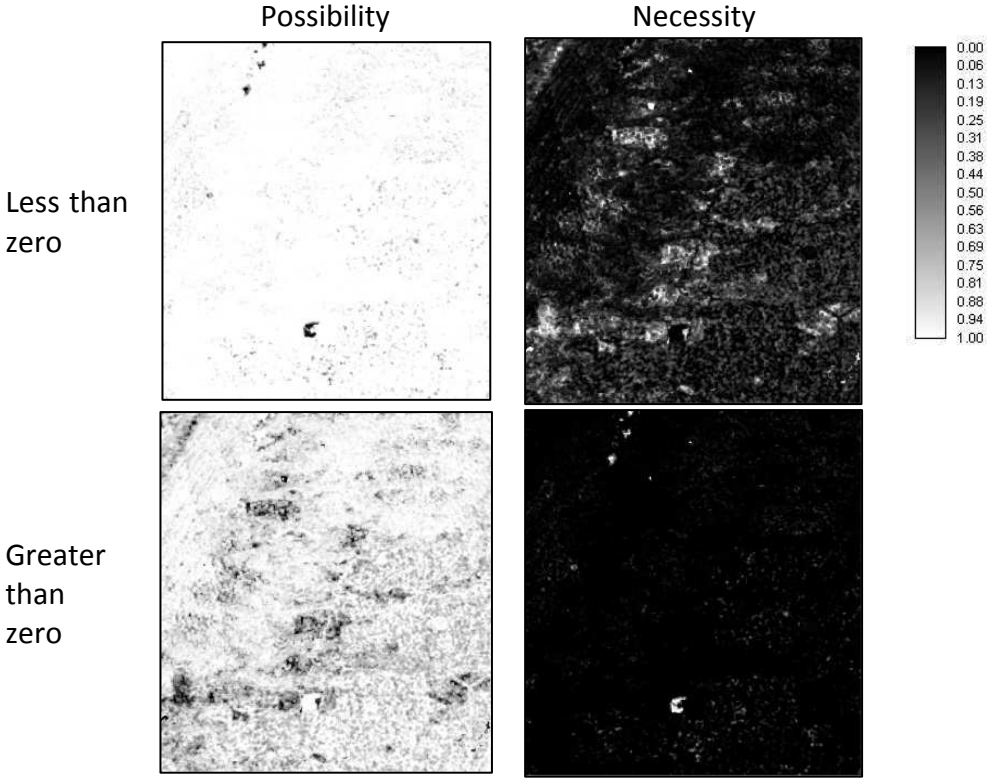



Figure 5: Possibility and necessity of each pixel being greater than or less than zero.

There are several possible combinations of possibility and necessity values that may be obtained from comparisons (Dubois and Prade 1983). If $\Pi_{\tilde{C}_{i,j}<(t)} = 0$ and $N_{\tilde{C}_{i,j}<(t)} = 0$, then the fuzzy number is definitely not smaller than the threshold $t$. If $\Pi_{\tilde{C}_{i,j}<(t)} > 0$ and $N_{\tilde{C}_{i,j}<(t)} = 0$ then it is possible that the fuzzy number is smaller but the indicators for this claim are not strong. Values of $\Pi_{\tilde{C}_{i,j}<(t)} = 1$, $N_{\tilde{C}_{i,j}<(t)} > 0$ provides a much stronger indication that the fuzzy value might be smaller than $t$. Only if $\Pi_{\tilde{C}_{i,j}<(t)} = 1$ and $N_{\tilde{C}_{i,j}<(t)} = 1$ can the fuzzy number be declared as definitely smaller than $t$. The interpretation of the values would be

similar for the inverse problem ($\tilde{C}_{i,j} > t$). For the area of the DSM shown in Figure 1, the possibility and necessity, that the change in elevation is greater than or less than zero are mapped in Fig. 5.

## 4. Conclusion

The proposed approach shows how the difference between surfaces can be calculated for two fuzzy surfaces and how the results can be identified as the change being greater than and less than some specific threshold (zero in the example presented here). The approach emphasizes uncertainty of the surface, not only in the calculation of difference but also in evaluation of the change which is achieved using two measures (possibility and necessity), that allow the capture of contradictory information. Contradictory areas may easily occur when dealing with uncertainty, because the values are no longer crisp and thus they cannot easily be categorized.

## Acknowledgements

The authors gratefully acknowledges support to the second author from the Operational Program for Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 and CZ.1.07/2.2.00/28.0078 of the Ministry of Education, Youth and Sports of the Czech Republic).

## References

Caha J, Vondráková A, Dvorský J, 2014, Comparison of Crisp, Fuzzy and Possibilistic Threshold in Spatial Queries. In: Abraham A, Krömer ., Snášel V (eds), Innovations in Bio-inspired Computing and Applications SE - 22. Berlin: Springer International Publishing, 239–248.

Dubois D, Prade H, 1983, Ranking Fuzzy Numbers in the Setting of Possibility Theory. Information Sciences, 30(3): 183–224.

Dubois D, Prade H, 1986, Possibility Theory: An approach to Computerized Processing of Uncertainty, New York: Plenum Press.

Fisher P F, Wood J, Cheng T, 2007, Higher order vagueness in a dynamic landscape: Multi-resolution morphometric analysis of a coastal dunefield. J. Environ. Inform. 9, 56-70.

Lodwick W, Anile A M and Spinella S, 2008. Introduction. In Lodwick W (ed), *Fuzzy surfaces in GIS and geographical analysis : theory, analytical methods, algorithms, and applications*, Boca Raton, CRC Press, 1–46.

Loquin K and Dubois D, 2010. Kriging with Ill-Known Variogram and Data. In Deshpande A and Hunter A (eds), *Scalable Uncertainty Management SE - 5*. Lecture Notes in Computer Science. Berlin, Springer, 219–235.

# Detecting errors in formally correct geodatabases

S. Savino, M. Rumor[1]

[1]IDepartment of Information Engineering, University of Padova, via Gradenigo 6, 35121 Padova
Email: { savinosa, rumor } @dei.unipd.it

## 1. Introduction

The development of digital geographic information systems revolutionized the world of cartography; among the many advantages to count, there is the possibility to query, transform and process the data, obtaining results that would not be practically achievable if still working on paper maps.

One of the hurdles to face while performing such operations is the presence of errors in the input.

Nowadays, performing more and more complex analysis on the data, errors that once were negligible have become unacceptable, as they hamper the soundness of our results: if digital data on the one hand allows for complex elaborations, on the other hand calls for good and correct input data.

For this purpose, most of modern GIS software have been fit with QA tools, able to detect errors in the data -and possibly correct them automatically- and it's now a common best practice while defining a new schema for a geodatabase, to define formal constraints both on spatial and non spatial attributes in the form of constraints on domain spaces and geometries, and attribute and topologic relations to reduce the risk of populating the data with incorrect values.

These tools and techniques allow to avoid, or to detect and correct, many errors on the input data and enable to create, eventually, a database that is formally correct, i.e. it respects all the constraints formalized on it.

Nevertheless, it's a common experience that also a formally correct database can contain errors: as constraints are able to formalize only the "structure" of the data, errors on its content can not be detected: an example of these are classification errors.

This kind of errors are very subtle, as they usually go undetected by common QA techniques and they remain unnoticed unless they happen to be detected when some results do not meet the expectations.

This paper will show the first results of a research project started in the GIS lab of the Department of Information Engineering of the University of Padova aimed at investigating and developing techniques for the quality analysis of geodatabases that focuses on errors that can not be detected by standard QA tools.

To go around the limits of current QA tools, the first step of the research was to move further from the traditional idea of error introducing the concept of "anomaly".

### 1.1 The concept of “anomaly”

To introduce what we mean with “anomaly”, it is better to use an example.
Let’s say that analyzing a dataset, a building object whose type attribute reads “lighthouse” is found located in the middle of a medium-sized town, far away from every water body. This lighthouse does not violate any constraint: its geometry is valid, it abides every topological constraint set on it and its attributes have values belonging to their respective domains spaces.

This object is formally correct. Nevertheless, according to common-sense, we could agree that it's probably flawed by a classification error, as it is quite unusual –to say the least– for a lighthouse to sit amid a town far away from the water.
This specific lighthouse is what we intend as an anomaly.

An anomaly can be defined as: an object that conveys information that does not conform to the expected characteristics of that particular object.
The basic idea of our approach is that by searching the data for values that are not within the range in which lie what we can accept as "good values", we are able to identify potential errors that have not been detected by the standard QA tools; these "not conforming" data are signalled to the user, who will evaluate them and decide whether they represent actual errors or not.

In the following sections a classification of anomalies is given first, followed by some first test results.

## 2. A taxonomy for anomalies in geodatabases

It's important to underline that an anomaly is different from a "formal" error: an anomaly does not violates any constraint formalized on the data; the latter, on the contrary, does, and can be detected by standard QA tools.
On the other hand, an anomaly is also different from an error, as it is not true that every anomaly corresponds to an error: an anomaly is just a hint that the information stored in the geodatabase might be incorrect; the final evaluation must be done by a human expert.

We first found three main aspects which characterize an anomaly and finally identified five types of anomaly; we also found that in the case of networks, because of their special properties, specific aspects need to be identified. The list is the following:

- shape
  - regularity
  - semantic dimension
- cardinality
  - respect to total number
  - respect to distribution
- position
  - respect to other features
- network
  - classification
  - connectivity
  - direction

### Shape anomalies

Regularity: according to expectations , man-made objects have a shape that is mostly regular, while not artificial objects have irregular shapes. With respect to this idea, a squared object is most likely to be a building and not a lake, and a regular shaped lake is an anomaly.

Semantic dimension: according to common sense, some objects have some minimum or maximum dimension (i.e. area and length); with regard of this, a road whose length is 1 meter or a wood patch whose area is 10 square meters represent an anomaly.

### Cardinality anomalies

Number of items: if the number of features and their semantic value can be related, counting them can be used to detect anomalies: for example finding an excessive number of parking lots in a remote area means they are anomalies.

Distribution: analyzing the way the number of features is spatially distributed on the map, it is possible to detect anomalies: the presence of churches in a city is supposed to be evenly distributed, while industrial silos are supposed to be clustered into groups; a distribution of these objects not obeying these rules means they are anomalies.

### Position anomalies

Relative position: as in the lighthouse example, there are many objects that, according to expectations, are spatially related to other objects; a lighthouse far from the sea is an anomaly, as it is a control tower far from a runway or a tower bell far from a church or a mountain hut close to the sea.

### Network anomalies

Classification
As a network represents connected elements and connected elements are spatially related in the real world, the characteristics of these object, and hence their classification, should not change abruptly but should be similar for elements close to each other: a road network whose edges change classification too often indicates they are anomalies.

Connectivity
A network is supposed to be mostly connected: a network having too many isolated edges means they are anomalies. When evaluating the connectivity of a network, also the dimension should be considered, as an edge or a gap in the network are supposed to have sensible lengths and, if they are too small, they are anomalies.

Direction
If a network is made of directed edges, their orientation could identify anomalies: for example a road from which you can only exit or a river flowing uphill are anomalies.

On the base of this classification, we developed some first tools to detect anomalies and we tested them on a sample dataset; the first results obtained are illustrated in the next section.

## 3. Detecting anomalies: first results

The algorithms to detect anomalies have been developed in Java as a plugin of OpenJUMP and tested on a sample dataset containing 1:2000 scale data of the Regione Veneto.

As the detection of anomalies relies on detecting data having values that “look strange”, these algorithms are based on some assumptions on what correct data should look like with respect to the classification given before. The rules to detect anomalies and the thresholds used have been empirically set.

Following are some results: the figures represent data from the test dataset flagged as anomalies, while the caption describes the anomaly detected.

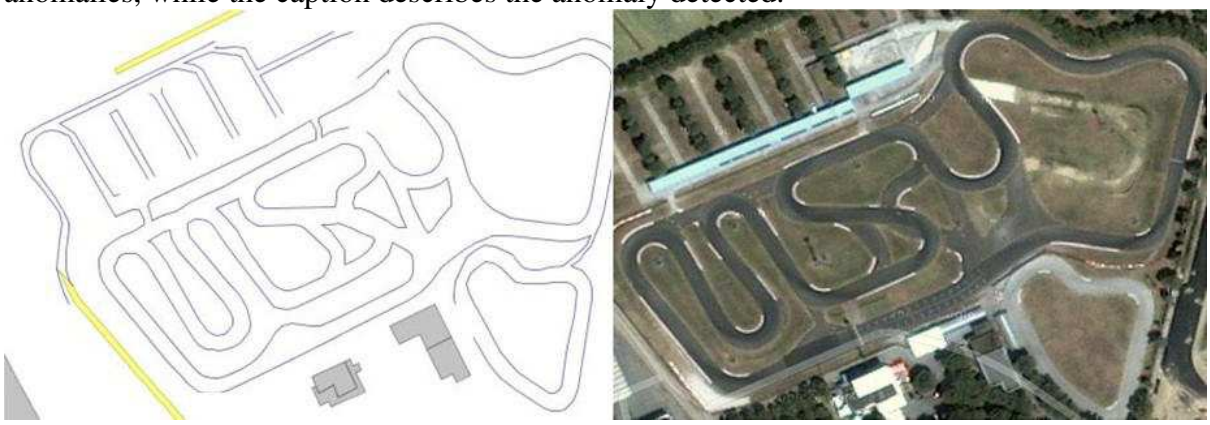


Figure 1. Shape regularity anomaly: the man made feature on the left is too irregular; it actually is a go-kart track.

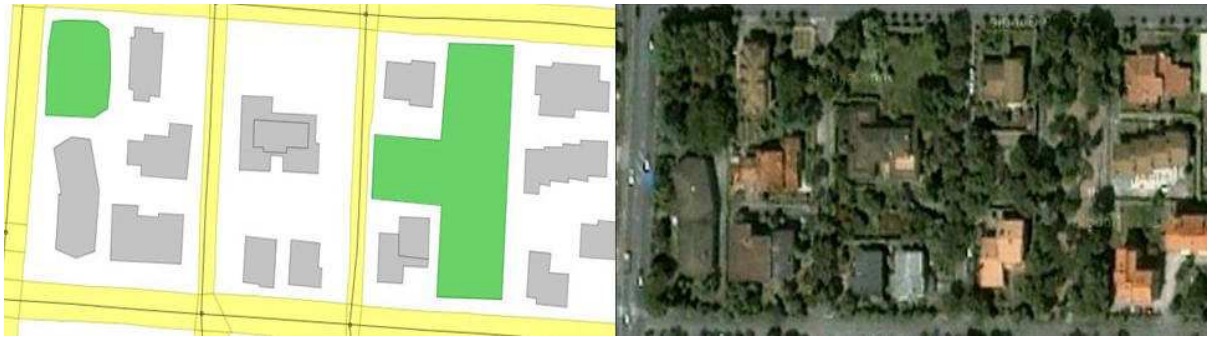


Figure 2. Wood patches are shown in green; the bigger is too regular for a natural feature (shape regularity anomaly), the smaller is too small to be a wood patch (semantic dimension anomaly); they should have been classified as urban green areas.

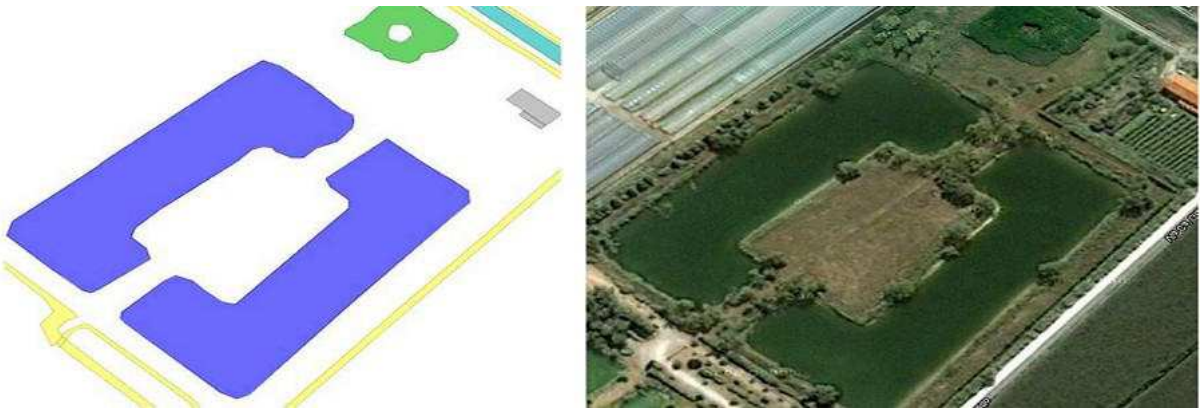


Figure 3. Shape regularity anomaly: the lake in blue is too regular. It actually is a game fishing pond.

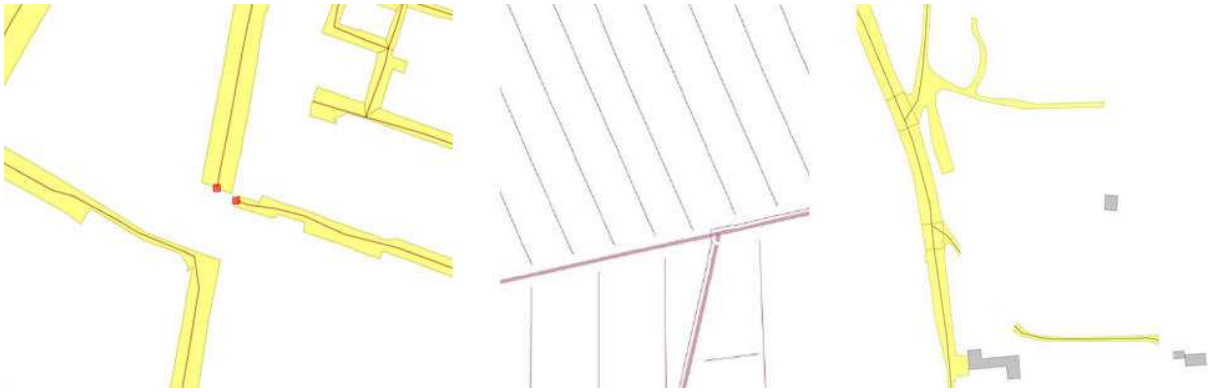


Figure 4. Network anomalies: the presence of small gaps (left and center image) and of an isolated edge (right image) are anomalies.

## 4. Conclusions

Despite modern QA tools and the possibility to define constraints on both geometric and semantic attributes of a geodatabase, data is still susceptible to errors that can not be detected, for example classification errors or position errors.
However, it is possible to develop tools that are able to detect anomalies in the content of a geodatabase: pieces of information that do not conform with an expected "good value".

Although anomalies per se are not errors, finding them allows the identification of errors that would be not detectable otherwise.

These tools are based on a cross-analysis of both spatial and semantic data and a comparison with expected values.

Using anomalies requires a shift in the usual approach to quality control: if using constraints means to define how wrong data looks like, to detect anomalies we first need to define how good data look like.
This requires to define many rules and thresholds; although the given classification helps their definition, they can not be general, but need to be customized on the basis of the information stored in the database (i.e. its schema) and of the characteristics of the related area. It is a long and not trivial task, that is probably best suited for very large datasets, exactly those for which a manual validation process is not feasible.

## Acknowledgements

The authors would like to acknowledge the work of Lorenzo Valerio and Alan Stocco, students of the Master Degree and Bachelor Degree of the Department of Information Engineering of the University of Padova who contributed during their thesis to the results of this research project.

## References

Louwsma J, Zlatanova S, Lammeren Rv, Oosterom Pv, 2006, Specifications and implementations of constraints in GIS- with Examples form a Geo-Virtual Reality System, GeoInformatica, 10:531-550.

Mas S and Reinhardt W, 2009, Categories of Geospatial and Temporal Integrity Constraints, Proceedings of the International Conference on Advanced Geographic Information Systems & Web Services, Cancun, Mexico, 146-151.

McCain M, 2003, Geodatabase Quality Control, now more important than ever, Proceedings of the 23rd Annual ESRI International User Conference, San Diego, California, 7-11.

Servigne S, Ubeda T, Puricelli A, Laurini R, 2000, A Methodology for Spatial Consistency Improvement of Geographic Databases, GeoInformatica, 4:7-34.

Shi W, Fisher P and Goodchild MF, 2003, Spatial data quality, CRC Press.

Udagepola KP, Xiang L, Xiaozong Y, Wijeratne AW, 2006, Review of data consistency and integrity constraints in spatial databases, Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, 348-353.

# A Time-Geographic Framework for Computing Spatio-Temporal Interaction Probabilities

J. A. Downs[1], D. Lamb[1], G. Hyzer[1], R. Loraamm[1], Z. J. Smith[1], and B. M. O'Neal[1].

[1]School of Geosciences, University of South Florida, Tampa, FL, USA
Email: downs@usf.edu, {davidlamb, hyzer, rloraamm, zsmith, bmoneal}@mail.usf.edu

## 1. Introduction

Quantifying interactions between mobile objects is a common task in GIScience. Interactions are commonly quantified from tracking data in three main ways: (1) rates of co-location (Miller 2012), (2) percentage of home range overlap (Olsen et al. 2011), and (3) measures of space-time prism intersection (Kwan 2000). Rates of co-location, while valuable, can only be computed if objects are tracked with near identical sampling schemes. Home ranges delineate the physical area occupied by individuals (often animals), and are often estimated using hull-based or statistical methods (Downs and Horner 2009, Downs et al. 2011, Worton 1987). While the amount of home range overlap can measure the potential spatial interaction of individuals, this approach does not in any way measure temporal aspects of interaction. Space-time prisms from time geography can be used to map the potential spatial locations objects over time given known spatial, temporal, and physical constraints that limit their movements (Hägerstrand 1970, Miller 2005a). When space-time prisms are mapped for multiple individuals, the potential for interactions can be measured both spatially and temporally by intersecting the prisms (Miller 2005b, Neutens et al. 2007). A limitation of this method is that it only measures potential for interaction. This paper extends that approach by calculating the probability of interaction. Specifically, voxel-based probabilistic space-time prisms are used quantify interaction probabilities for objects in space and time.

## 2. Voxel-based Probabilistic Space-Time Prisms

The concept of a probabilistic space-time prism was introduced by Winter and Yin (2010b, 2010a), while Song and Miller (Song and Miller 2013) provided an additional theoretical basis for their derivation. Probabilistic space-time prisms extend the traditional space-time prism to map the probability—rather than possibility—of an object's movement through space and time. Downs et al. (2013) developed a voxel-based approach for the geocomputation of probabilistic space-time prisms, which is used here. Traditional space-time prisms can be generalized using voxels in three-dimensional space-time. Here, each individual voxel represents a spatial area at a specific time interval. Once voxels in a study area are defined based on the desired spatial and temporal resolution, a space-time prism can be computed by determining which spatial locations are accessible to the object at each time given a set of control points and its maximum velocity. Each voxel is encoded with a 1 or 0 to indicate whether it is contained in the prism or not. Mathematically, a space-time prism (STP) can be formulated for each voxel, denoted $l_{ba}$, using the following equation, as per Downs et al. (Downs et al. 2013):

$$STP_{l_{b,a}} = \begin{cases} 1, \text{ if } \|x_a - x_i\| \leq (t_b - t_i)v \text{ and } \|x_j - x_a\| \leq (t_j - t_b)v \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

where:

$C = \{c_1(t_1, x_1), \ldots, c_i(t_i, x_i), c_j(t_j, x_j), \ldots, c_n(t_n, x_n)\}$ is the set of $n$ tracking data points, where each control point $c$ is indexed as $i$ with timestamp $t_i$ and spatial location $x_i$. Control points immediately following $c_i$ are denoted $c_j$ with timestamp $t_j$ and location $x_j$

$R = \{r_1(x_1), \ldots, r_a(x_a), \ldots, r_m(x_m)\}$ is the set of $m$ raster cells $r$ in the study region, where each raster cell indexed as $a$, where raster cell $r_a$ has spatial location $x_a$ recorded at its centroid.

$K = \{k_1(t_1), \ldots, k_b(t_b), \ldots, k_q(t_q)\}$ is the set of $q$ time intervals $k$ indexed as $b$, where timestamp $t_b$ is recorded at the midpoint of the time interval $k_b$.

$L = \{l_{11}(k_1, r_1), \ldots, l_{ba}(k_b, r_a), \ldots, l_{qm}(k_q, r_m)\}$ is the set of voxels $l$ that contains the space-time prism indexed as $ba$, where voxel $l_{ba}$ corresponds to raster cell $r_a$ at time interval $t_b$. $L_k$ is used to denote the subset of voxels for a particular time interval, or space-time disk. $L_r$ is used to denote the collection of voxels for a spatial location.

$\| \cdot \|$ calculates the distance between two spatial locations $x$

$v$ defines the maximum velocity of the object.

A voxel is contained in the STP between two consecutive points if both these conditions are true: (1) the distance between the first tracking point and the voxel is less than or equal to the maximum distance the object could have moved between those locations and (2) the distance between the voxel and the next tracking point is less than or equal to the maximum distance the object could have moved between those locations. The maximum movement distance is calculated based on the amount of time elapsed and the estimated velocity of the object. If either of these measured distances is greater, then the voxel must be located outside the STP, in which case it is assigned a 0.

Once a space-time prism is created using binary-encoded voxels, probabilities can be derived by using a distance-weighting function to distribute probabilities within each space-time disk. For instance, an inverse distance-weighting formula can be used to obtain the probabilities, where the probability at a particular voxel, $P(STP_{l_{b,a}})$, is computed as:

$$P(STP_{l_{b,a}}) = \frac{\frac{1}{\|x_s - x_a\|}}{\sum_{l_{b,a} \in L_k} \frac{1}{\|x_s - x_a\|}} \quad (2)$$

where $\|\boldsymbol{x_s} - \boldsymbol{x_a}\|$ measures the distance between the voxel's spatial location and the location of estimated space-time path at time k. In this case, the probability is computed by first calculating 1 divided by the distance between the voxel and the space-time path and then dividing that value by the sum of all distance-weighted values for all voxels in the same space-time disk. All probabilities for a given space-time disk sum to 1. Further details about the geocomputational procedure can be found in Downs et al. (2013) Figure 1 illustrates sample voxel-based probabilistic space-time prisms for two individuals, where darker colours indicate higher probabilities.

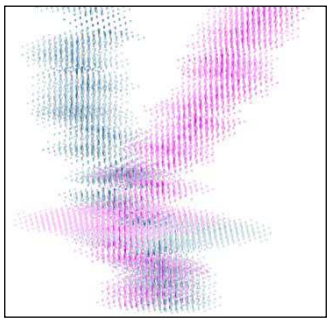
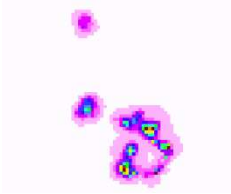


Figure 1. Sample probabilistic-space time prisms, symbolized using voxel-centroids.

## 3. Computation of Interaction Probabilities

Once voxel-based space-time prisms are computed for multiple individuals, they can be intersected to derive a number of probabilities. This research proposes geocomputational and geovisualization procedures for three types of probabilities. Probability equations and geovisualization approaches are summarized in Table 1. All calculations assume that independent individuals are labelled A, B, C, …, Z and are formulated using notation from section 2.

Table 1. Spatio-temporal interaction probability equations and geovisualization techniques.

| Probability | Formula | Geovisualization |
|---|---|---|
| Independent objects were together at a specific location at a specific time | $P(A \cap B \cap \ldots \cap Z)_{l_{b,a}} = P(A)_{l_{b,a}} \times P(B)_{l_{b,a}} \times \ldots \times P(Z)_{l_{b,a}}$ | Calculated for each voxel; symbolized by time step (2D view below) or stacked for multiple time steps (3D)  |
| Independent objects were together at a specific time (at any location) | $P(A \cap B \cap \ldots \cap Z)_{t_b} = [P(A \cap B \cap \ldots \cap Z)_{l_{b,1}}] \cup [P(A \cap B \cap \ldots \cap Z)_{l_{b,2}}] \cup \ldots \cup [P(A \cap B \cap \ldots \cap Z)_{l_{b,m}}]$. | Calculated for each time step; multiple time steps can be geovisualized using probability clocks (below)  |
| Independent objects were together at a specific location (at any time) | $P(A \cap B \cap \ldots \cap Z)_{r_a} = [P(A \cap B \cap \ldots \cap Z)_{l_{1,a}}] \cup [P(A \cap B \cap \ldots \cap Z)_{l_{2,a}}] \cup \ldots \cup [P(A \cap B \cap \ldots \cap Z)_{l_{q,a}}]$ | Calculated for each raster cell; results for all cells can be geovisualized as a probability map (below)  |

## Acknowledgements

Portions of this research were funded by grants made to the lead author (Downs) from the National Science Foundation (NSF) [grant number BCS-1062947]. The contents of this article are the responsibility of the authors and do not reflect the views of the NSF. The research was also supported by the University of South Florida (USF) College of Arts and Sciences Internal Awards Program.

## References

Downs J, Horner M, 2009, A characteristic-hull based method for home range estimation. Transactions in GIS 13:527-537.

Downs J, Horner M, Hyzer G, Lamb D, and Loraamm R, 2013, Voxel-based probabilistic space-time prisms for analysing animal movements and habitat use. International Journal of Geographical Information Science 28 (5):875-890.

Downs J, Horner M, and Tucker A, 2011, Time-geographic density estimation for home range analysis. Annals of GIS 17 (3):163-171.

Hägerstrand T, 1970, What about people in regional science? Papers of the Regional Science Association 24:7-21.

Kwan M-P, 2000, Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. Transportation Research Part C: Emerging Technologies 8 (1-6):185-203.

Miller H, 2005a, A measurement theory for time geography. Geographical Analysis 37 (1):17-45.

Miller H, 2005b, Necessary space-time conditions for human interaction. Environment and Planning B-Planning & Design 32 (3):381-401.

Miller, J, 2012, Using spatially explicit simulated data to analyze animal interactions: a case study with brown hyenas in Northern Botswana. Transactions in GIS 16 (3):271-291.

Neutens T, Witlox F, De Weghe N, and De Maeyer P, 2007, Human interaction spaces under uncertainty. Transportation Research Record (2021):28-35.

Olsen J, Downs J, Tucker A, Trost S, 2011, Home-range size and territorial calling of southern boobooks (ninox novaeseelandiae) in adjacent territories. Journal of Raptor Research 45 (2):136-142.

Song Y and Miller H, 2013, Simulating visit probability distributions within planar space-time prisms. International Journal of Geographical Information Science 28 (1):104-125.

Winter S and Yin Z, 2010a, Directed movements in probabilistic time geography: Taylor & Francis, 1349 - 1365.

Winter S and Yin Z, 2010b, The elements of probabilistic time geography GeoInformatica DOI: 10.1007/s10707-010-0108-1.

Worton B, 1987, A review of models of home range for animal movement. Ecological Modelling 38 (3-4):277-298.

# Coerced Geographic Information: The Not-so-voluntary Side of User-generated Geo-content

Grant McKenzie, Krzysztof Janowicz

Department of Geography, The University of California, Santa Barbara, USA
Email: {grant.mckenzie; jano}@geog.ucsb.edu

## 1 User-generated Geo-content

It has been seven years since Michael Goodchild published his widely visible paper describing citizens as sensors and defining user-generated content of a geographic nature as *Volunteered Geographic Information (VGI)* (Goodchild, 2007). Two years prior to this publication, the term *User-generated Content (UGC)* began to appear in the media alongside such terms as *Web 2.0*, *Social Web*, and *New Media*. In the past seven years, the world of user-contributed content has changed substantially. While individuals and groups still actively contribute geographic information to open platforms such as *OpenStreetMap* and *Wikimapia*, an increasing trend has been to contribute data, a lot of which contains geographic attributes, to private data silos. In many cases the motivation for contributing to these private data sources is completely unrelated, or at least secondary, to the actual data being captured. Unbeknownst to many users of these systems, their data are being used for purposes other than those originally intended, both by commercial entities as well as the academic research community. Furthermore, the ability to access data contributed to these systems deviates significantly from the open-access nature of traditional VGI systems. The modern private silos are largely *data one-way-streets*.

In this short paper we explore the differences between what has historically been labeled VGI and what we propose should lie under the title of *Coerced Geographic Information (CoGI)*. We give examples of datasets that are generated through different means and outline a set of five criteria that may be used to define the differences between VGI and CoGI. Lastly, we present a *VGI to CoGI Scale* that is used to rate current platforms that collect user-generated geo-content.

The value of geospatial data has not gone unnoticed in the recent rise of online social networking (OSN). Virtually every social media platform established in the last five years either began with or has come to incorporate geographic data in their applications. Platforms such as *Foursquare*, *Yelp* and *Jiepang* were founded in geospatial data offering users the ability to *check-in* to places or share their mobile device location with friends. More popular applications such as *Facebook*, *Twitter* and even *Sina Weibo* began without the ability to geotag content and later added this feature as *(a)* geolocation technology became ubiquitous and *(b)* the companies realized the power in knowing the locations of their users. While the ability to geotag photos or check-in to a specific location has been

sold as a *feature* to users of these platforms, it has come at the cost of location privacy, a concept that is not fully explained by the platforms nor understood by the average user contributing to these applications.

Take for example a user uploading photos to *Yahoo's* photo sharing application *Flickr*. The primary purpose of uploading photos, to most users, is to share moments with their friends, family or even the public. The fact that many mobile image capturing devices (e.g., mobile phones, digital cameras) include a location tag in the image header is either not known to the original publisher or seen as a secondary *feature* of arguably little importance. What is truthfully not understood by the vast majority of contributors is the value of this "secondary" geodata. Previous work (Girardin et al., 2007; Toyama et al., 2003) has shown just how rich this geodata is through the construction of gazetteers and travelogues. The *Flickr* API even offers the ability for developers to extract the location information directly from a photograph's exchange image file format (Exif).[1] If the academic community is able to construct such robust data models with a minimal amount of data accessed through public-facing APIs, one can only imagine what is possible given the full set of data.

A second example is the gamification aspects of online social networks. Applications such as *Foursquare* offer users the ability to check-in to places in order to gain points and receive *badges*. The more places you check-in, the more points you receive. From a social perspective this is quite appealing. I can compete with my friends for points and *mayorships*. The *only* cost of this entertainment is sharing my location. While the game-play and enjoyment of users is the motivating factor, the benefit to Foursquare is enormous. The company has built an entire business model around geospatial data contributed by their users. At last count, Foursquare lists more than 50 million[2] *venues* (Points of Interest) contributed by more than 50 million users. These sensitive personal data are used for location-targeted advertising, business registration, and is even sold to third parties (Van Grove, 2013).

OSN contributions aside, location data is being gathered from a great range of sources. Credit card companies have access to location-specific transaction records and reward programs keep track of where people buy fuel and groceries. Opening the *Privacy Settings* on a location-enabled mobile device will present the user with all applications that have access to location data. Applications such as a *Brightest Flashlight* or *Angry Birds* have no motivation for collecting location information other than resale (Hong, 2012). One of the primary question of interest is if users are aware of what is being collected and done with their data.

From a research community perspective it is important to realize that many individuals contributing data to platforms such as *Twitter* are unaware that their data are being used for research purposes. A limited few may be aware that their location data is used for direct marketing and advertising, but the vast majority of contributors are unaware that existing research employs psychological profiling techniques as well as terrorist threat detection models (Mahmood, 2013), for example, on their seemingly private data and that their tweets may appear as examples in scientific papers.

[1] https://www.flickr.com/services/api/flickr.photos.getExif.html

[2] These are speculative numbers as Foursquare keeps this information private

## 2 Criteria and Scale

Given the examples listed above it is clear that there should be some criteria that can be applied to any given user-contribution platform in order to place it on a VGI to CoGI scale. To start the discussion, we introduce five criteria in this section each of which can be used to rate a platform on a 3 point scale from **0** (highly CoGI) to **2** (highly VGI).

(I) **Equivalent bi-directional data access**. One of the truly limiting aspects of CoGI and one that explicitly differentiates it from traditional VGI is the accessibility of the contributed data. The value of traditional VGI platforms is that it is as easy to access the data as it is to contribute. *OpenStreetMap*, for example allow users to download entire *planet* files while platforms such as *Twitter* restrict consumption of data through limited APIs and application such as *Strut* offer no ability to consume the contributed data, i.e., they are data one-way streets.

(II) **Limited terms of use**. This criteria relates to the restrictions on how the data can be used once consumed. Conventional VGI systems, e.g., Wikimapia, have very open (often *free to all*) licenses, while more commercially oriented systems often claim the rights to contributed data and restrict the terms of use. Still other platforms act as data silos, allowing no external reuse of the data.

(III) **Awareness of Contributed Data**. An important question to ask of any platform is whether the users are aware of the data that they are contributing. For example users may not be aware that their IP address is being recorded or that their location may be inferred through Wi-Fi positioning. CoGI platforms range in their level of transparency regarding what is being collected.

(IV) **Awareness of Data Usage**. Transparency on how contributed data is used is an important criteria. Sites like *OpenStreetMap* offer a reasonable level of transparency when it comes to data usage. Most contributors realize that the data they contribute can be used for almost any purpose. In contrast it is less likely that users of the *Foursquare* application are aware that their *check-ins* are being used to target advertising at their friends or even for predictive policing.

(V) **Active User Involvement**. Lastly, *active* vs. *passive* user involvement is of interest. The act of downloading an application and contributing data to it indicates active involvement in the user-generated content process. Alternatively, geosensors such as bluetooth, RFID tags or CCTV cameras do not offer users the option of contributing data but rather collect content generated by users. Existing research in this area has focused on related concepts of *opt-in* versus *opt-out* provisions as they relate to crowd-sourced data (Harvey, 2013). A particularly malicious example involves trash bins in the city of London that spy on the MAC addresses of mobile phones to determine the location and movement pattern of citizens.[3]

Given the above criteria it is possible to describe user-generated geo-content platforms in terms of their VGI-CoGI*ness*. Table 1 shows a sample of six such platforms along with a 0-2 scalar rating for each criterion (high CoGI to high VGI respectively). The total is calculated across all criteria with high values, maximum 10, indicating a high tendency towards VGI and low values, minimum 0, depicting applications leaning towards CoGI.

[3] http://www.bloomberg.com/news/2013-08-12/snooping-garbage-bins-in-city-of-london-ordered-to-be-disabled.html

Table 1. Scalar of sample platforms based on five criteria. A value of "0" indicates highly CoGI while a value of 2 represents highly VGI (per category).

| Platform | I | II | III | IV | V | Total |
|---|---|---|---|---|---|---|
| OpenStreetMap | 2 | 2 | 1 | 2 | 2 | **9** |
| Flickr | 1 | 1 | 1 | 1 | 2 | **6** |
| WikiMapia | 2 | 2 | 1 | 2 | 2 | **9** |
| Foursquare | 1 | 1 | 1 | 0 | 1 | **4** |
| Google Map Maker | 1 | 1 | 1 | 1 | 2 | **6** |
| Brightest Flashlight App | 0 | 0 | 0 | 0 | 1 | **1** |

## 3 Conclusions

The recent increase in user-generated geo-content has given rise to a vast number of platforms eager to provide tools for data contribution as well as data consumption. While methods for generating truly *Volunteered Geographic Information* have continued to thrive in this environment, a new division of user-generated geo-content has emerged, one that is not-so-volunteered. As commercial entities have realized the power of user-level geographic information, they have developed tools to ascertain this information. In most cases it is not clear what data are collected, how they will be used, who has access to these data (internally), to whom the data may be sold (externally), and to which degree the data are integrated with other sources of personal information. This paper outlines a number of ways in which VGI and CoGI differ and presents important distinctions of which users of user-generated geodata should be aware.

## References

Girardin, F., Dal Fiore, F., Blat, J., and Ratti, C. (2007). Understanding of tourist dynamics from explicitly disclosed location information. In *Symposium on LBS and Telecartography*. Citeseer.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Harvey, F. (2013). To volunteer or to contribute locational information? towards truth in labeling for crowdsourced geographic information. In *Crowdsourcing Geographic Knowledge*, pages 31–42. Springer.

Hong, J. (2012). Analysis of most unexpected permissions for android apps. Online. http://confabulator.blogspot.com/2012/11/analysis-of-top-10-most-unexpected.html.

Mahmood, S. (2013). Online social networks and terrorism: Threats and defenses. In *Security and Privacy Preserving in Social Networks*, pages 73–94. Springer.

Toyama, K., Logan, R., and Roseway, A. (2003). Geographic location tags on digital images. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 156–166. ACM.

Van Grove, J. (2013). Foursquare partners with gnip to sell check-in stream. Online. http://www.cnet.com/news/foursquare-partners-with-gnip-to-sell-check-in-stream.

# Modeling Visit Probabilities within Network Time Prisms using Markov Techniques

Ying Song[1], Harvey J. Miller[1], Xuesong Zhou[2]

[1]Department of Geography, The Ohio State University,
1036 Derby Hall, 154 North Oval Mall, Columbus, OH 43210, U.S.A.
Email: {song.851, miller.81}@osu.edu

[2]School of Sustainable Engineering and the Built Environment, Arizona State University,
Engineering G-Wing, 501 E Tyler Mall, Tempe, AZ 85287, U.S.A.
Email: xzhou74@asu.edu

## 1. Introduction

A core technique for measuring potential mobility is the *space-time prism* (STP): the envelope of all possible space-time paths between two known locations and times. However, the prism in its classic form is binary: all locations within the prism are considered to be equally accessible. Recent investigations into the planar STP suggest that this masks interesting properties of the prism interior. In particular, the distribution of visit probabilities within a STP is unequal: locations towards the center are more likely to be visited since there are more possible paths relative to locations near the boundary (Winter and Yin 2010a, 2010b, Song and Miller 2013).

This research develops methods for modeling visit probabilities within *network time prisms* (NTPs), that is, prisms constrained by transportation networks (Kuijpers and Othman 2009; Miller 1991). We adapt Markov techniques according to two basic types of mobility: i) non-vehicle movement where speed and direction are loosely constrained by the network (e.g. walking); and, ii) vehicle-based mobility where speed and direction are tightly constrained by the network (e.g., bicycles, automobiles).

## 2. Modeling Visit Probabilities within Network Time Prisms

We represent a spatial network as a graph $G = (V, E)$, with road intersections as a set of vertices $V = \{v_1, v_2, v_3, ...\}$ and road segments as a set of edges $E = \{e_{ij}\}$ with travel time between two adjacent vertices $t_{ij} = t_{SN}(v_i, v_j)$.

### 2.1 Visit Probabilities for Non-vehicular Movement

The probability of visiting an edge $e_{ij}$ at $t \in [t_O, t_D]$ is the joint probability of that edge being reached from the NTP origin anchor within $(t - t_O)$ and arriving at the destination anchor within $(t_D - t)$. For an undirected graph, we have $e_{ij} = e_{ji}$ and four types of movements along edges according to the entrance and exit vertices of each edge: i) $v_i \rightarrow v_j$; ii) $v_i \rightarrow v_i$; iii) $v_j \rightarrow v_i$; and iv) $v_j \rightarrow v_j$.

We use the earliest arrival time $t_i^-$ and latest departure time $t_i^+$ at a vertex $v_i$ to indicate potential space-time paths and define Brownian motion from the origin anchor and to the destination anchor respectively:

$$\varphi\left(B(t_i^-)\right) = \begin{cases} 0 & , \quad t_i^- - t_O < 0 \\ \dfrac{1}{\sqrt{2\pi} \times \delta \times (t - t_O)} e^{-\frac{(t_i^- - t_O)^2}{2\delta^2 (t - t_O)^2}}, & t_i^- - t_O \geq 0 \end{cases} \tag{1}$$

$$\varphi\left(B(t_i^+)\right) = \begin{cases} 0 & , \quad t_D - t_i^+ < 0 \\ \dfrac{1}{\sqrt{2\pi} \times \delta \times (t_D - t)} e^{-\frac{(t_D - t_i^+)^2}{2\delta^2 (t_D - t)^2}}, & t_D - t_i^+ \geq 0 \end{cases} \tag{2}$$

where $\delta$ is a dispersion parameter representing the level of mobility subject to the spatio-temporal constraints. We define the probability to move from $v_i$ to $v_j$ as:

$$P_{ij}(t) = \begin{cases} \varphi\left(B(t_i^-)\right) \times \varphi\left(B(t_j^+)\right) & , t_i^- \leq t \leq t_j^+ \\ 0 & , \text{otherwise} \end{cases} \tag{3}$$

The probability of moving along $e_{ij}$ is the summed probabilities of the four types of possible movements within the edge. We normalize the edge probabilities so that, for any time $t \in [t_O, t_D]$, they add up to unity:

$$P\left(e_{ij}, t\right) = \frac{\sum_{k=i,j,l=i,j} P_{kl}(t)}{\sum_{e_{mn} \in E} P(e_{mn}, t)} \tag{4}$$

## 2.2. Visit Probabilities for Vehicular Mobility

In vehicle-based mobility, speed limits, traffic directions, one-way and turn restrictions make it difficult for vehicles to move freely. Therefore, we adapt the continuous-time Markov process that accounts for both movement probabilities and transition rates. We modify this considering two factors. First, due to the speed limit, there is a minimum time required along each edge. Second, due to requirements to reach the destination on time, the conditional transition probabilities $\{p_{ij}\}$ are time-inhomogeneous and conditioned on the time $t \in [t_O, t_D]$.

We define a state as a vertex $v_i$; and a transition between states is a movement from $v_i$ to $v_j$ along an edge $e_{ij}$. We adopt and modify exponential distribution for holding time density function that describes the probability of arriving at $v_j$ at time $\tau$ after leaving $v_i$ at time $t$:

$$h_{ij}(\Delta\tau) = \begin{cases} \lambda e^{-\lambda \Delta\tau} & , \quad \Delta\tau \geq 0 \\ 0 & , \quad \Delta\tau < 0 \end{cases} \tag{5}$$

where $t_{ij} = t_{SN}(v_i, v_j)$ is minimum transition time for edge $e_{ij}$, $\Delta\tau = \tau - t_{ij}$ is extra time for transition, and $\lambda$ is the transition rate.

We derive the time profile for each edge $e_{ij}$, $(t_i^-, t_j^+, t_{ij})$ and use it to indicate potential arrival times and departure times. We consider a diffusion process from the origin: the

probability to pass a location within network depends on its available arrival times at $t \in [t_O, t_D]$: the arrival times follow the exponential distribution with rate $\lambda$. Therefore, the probability of reaching $e_{ij}$ from $v_i$ is:

$$p_i(t) = \begin{cases} 0 & , \quad t \in [t_O, t_i^-) \\ \int_{t_i^-}^{t} \lambda e^{-\lambda\tau} d\tau = e^{-\lambda t_i^-} - e^{-\lambda t} & , \quad t \in \left[t_i^-, t_j^+ - t_{ij}\right) \\ \int_{t_i^-}^{t_j^+ - t_{ij}} \lambda e^{-\lambda\tau} d\tau = e^{-\lambda t_i^-} - e^{-\lambda\left(t_j^+ - t_{ij}\right)} & , \quad t \in \left[t_j^+ - t_{ij}, t_D\right] \end{cases} \tag{6}$$

Similarly, we can derive the probability of reaching the destination from $v_j$ based on departure times:

$$p_j(t) = \begin{cases} 0 & , \quad t \in \left[t_j^+, t_D\right] \\ \int_{t_D - t_j^+}^{t_D - t} \lambda e^{-\lambda\tau} d\tau = e^{-\lambda(t_D - t_j^+)} - e^{-\lambda(t_D - t)} & , \quad t \in \left[t_i^- + t_{ij}, t_j^+\right) \\ \int_{t_D - t_j^+}^{t_D - t_j^+ - t_{ij}} \lambda e^{-\lambda\tau} d\tau = e^{-\lambda(t_D - t_j^+)} - e^{-\lambda(t_D - t_j^+ - t_{ij})} & , \quad t \in \left[t_O, t_i^- + t_{ij}\right) \end{cases} \tag{7}$$

The visit probability for $e_{ij}$ at $t$ is based on the time flexibility provided by both vertices, and normalized so that, for any $t \in [t_O, t_D]$, probabilities for state space add up to unity.

$$P\left(e_{ij}, t\right) = \frac{p_i(t) \times p_j(t)}{\sum_{e_{mn} \in E} P(e_{mn}, t)} \tag{8}$$

## 3. Results

We implement the method using Python and three modules that supported by Python: i) Numpy for creating and managing large-size arrays; ii) Scipy for basic statistic functions (e.g. probability density function for the normal distribution); and, iii) ArcPy for conducting network analysis and generating data to be visualized within ArcGIS.

Our analysis uses network and GPS data for New York City, USA. We chose New York City because of the availability of GPS data collected by a commercial navigation vendor. We chose Manhattan specifically since it allows us to use locations such as bridge and tunnel exits and entrances as surrogates for the NTP anchor points.

### 3.1. Scenario 1: Walking in Manhattan

Figure 1 illustrates NTP visit probabilities for walking between two locations in Manhattan with a 15 minute time budget. The planar STP potential path area is provided for reference. As can be

seen, the method generates movement probabilities that are intuitive and consistent with planar STP visit probabilities, albeit impacted by the constraints imposed by the network. (The NTP appears to extend beyond the boundary of the planar PPA at 11 minutes, but this is due to aggregating probabilities to the entire network edge.)

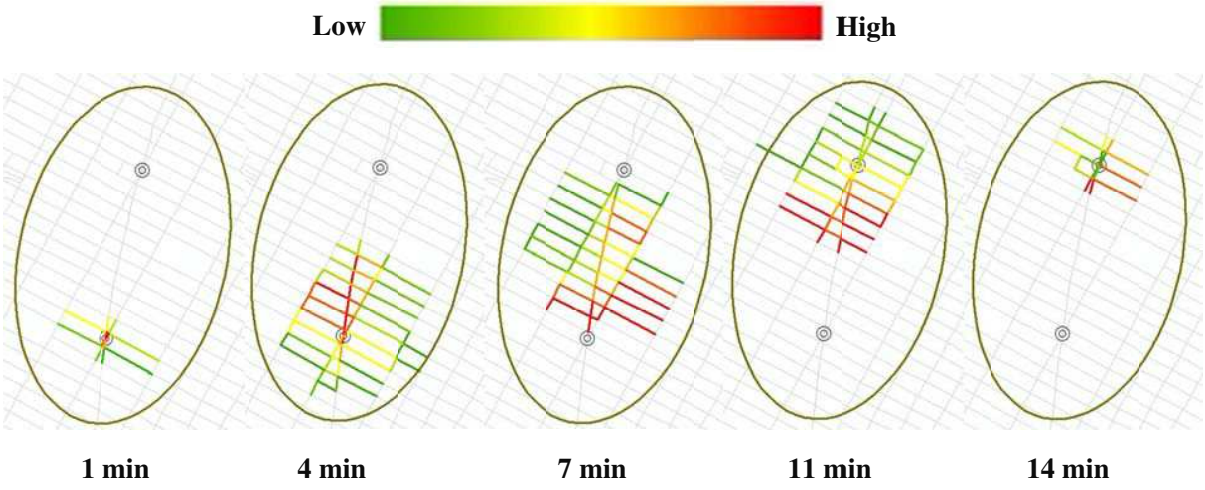



Figure 1. Visit probabilities aggregated by network edge.

### 3.2. Scenario 2: Driving across Manhattan

We calibrate the transition rate parameter $\lambda$ by selecting three routes that contain multiple sample trajectories within each NTP and using ordinary least squares estimation. All three sets of data indicate that the transition rate is 0.003 in the Manhattan area (with the time unit in seconds).

Figure 2 and 3 show the results for driving (1) across the southern Manhattan from the Holland Tunnel to the Manhattan Bridge within 25 minutes; and, (2) across the eastern Manhattan from the Third Avenue Bridge to the Queensborough Bridge (a.k.a. the 59$^{th}$ Street Bridge) within 40 minutes. The top half shows the simulated NTP visit probabilities while the bottom half shows the empirical probabilities derived from 72 and 104 GPS trajectories.

Table 1 provides results from calculating the root mean square errors (RMS) to quantitatively measure the difference between the simulated and empirical visit probability distributions. Two bases are used: (1) all feasible routes within the NTP that allows us to assess the accuracy of the simulated visit probabilities with respect to the empirical visit probabilities; and (2) routes used by empirical GPS traces that allow us to examine how well the simulated visit probabilities fit the empirical traces. Correspondence is good for both scenarios despite the small sample size, although some biases exist due to the commercial navigation vendor favoring primary streets over secondary streets in their routing guidance.

Table 1. Root mean square error between simulated and empirical distribution at selected time

(a) Southern Manhattan scenario

| Base | 5 min | 10 min | 15 min | 20 min |
|---|---|---|---|---|
| NTP | 3.0463% | 0.2646% | 0.4796% | 2.4739% |
| Empirical traces (n=72) | 1.0630% | 1.3892% | 2.5923% | 1.7000% |

(b) Eastern Manhattan scenario

| Base | 5 min | 10 min | 15 min | 20 min | 25 min | 30 min | 35 min |
|---|---|---|---|---|---|---|---|
| NTP | 0.5099% | 0.2106% | 0.2449% | 0.3017% | 0.3317% | 1.0954% | 6.5955% |
| Empirical traces (n=104) | 1.7292% | 0.9327% | 1.2014% | 1.4526% | 1.6371% | 2.9360% | 5.9691% |

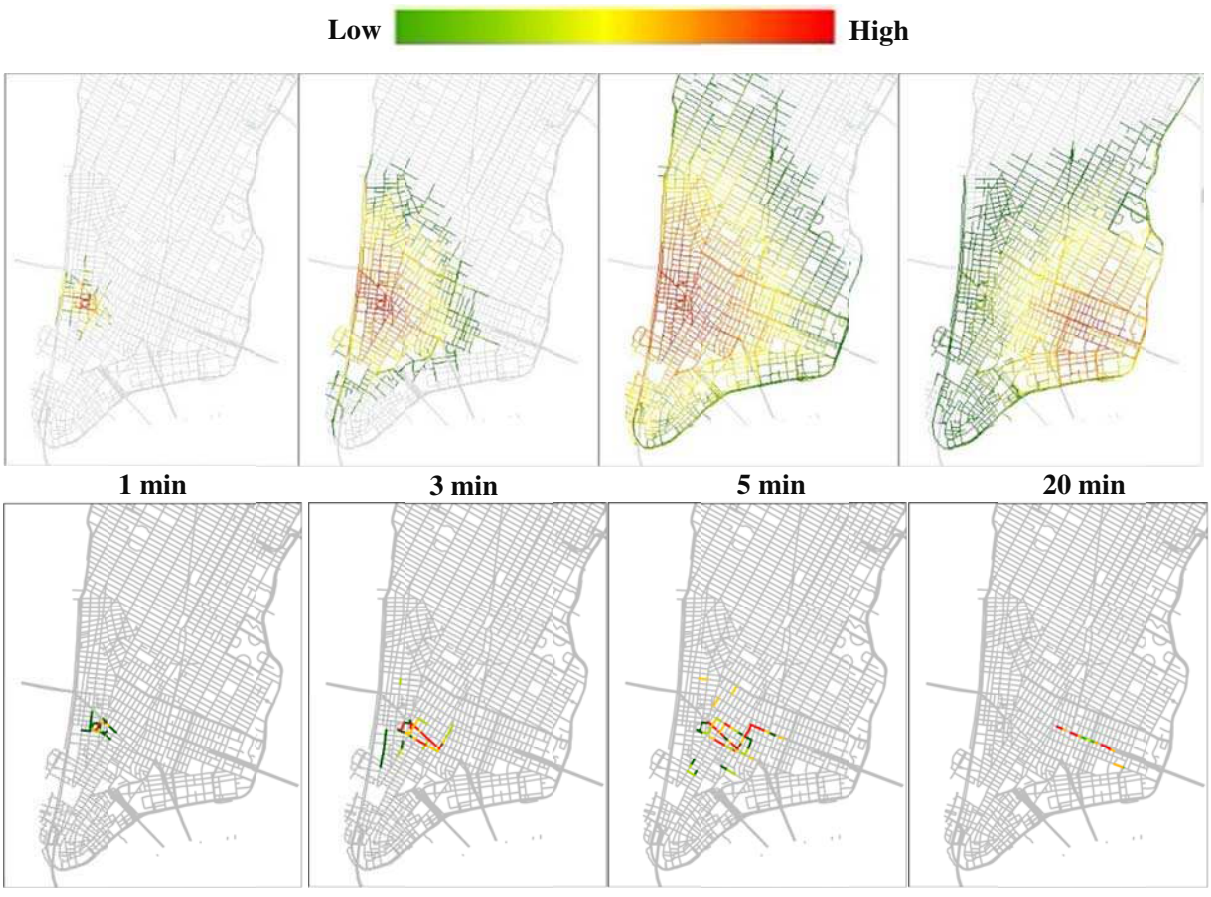



Figure 2. Comparison of simulated and empirical visit probabilities for the southern Manhattan scenario.

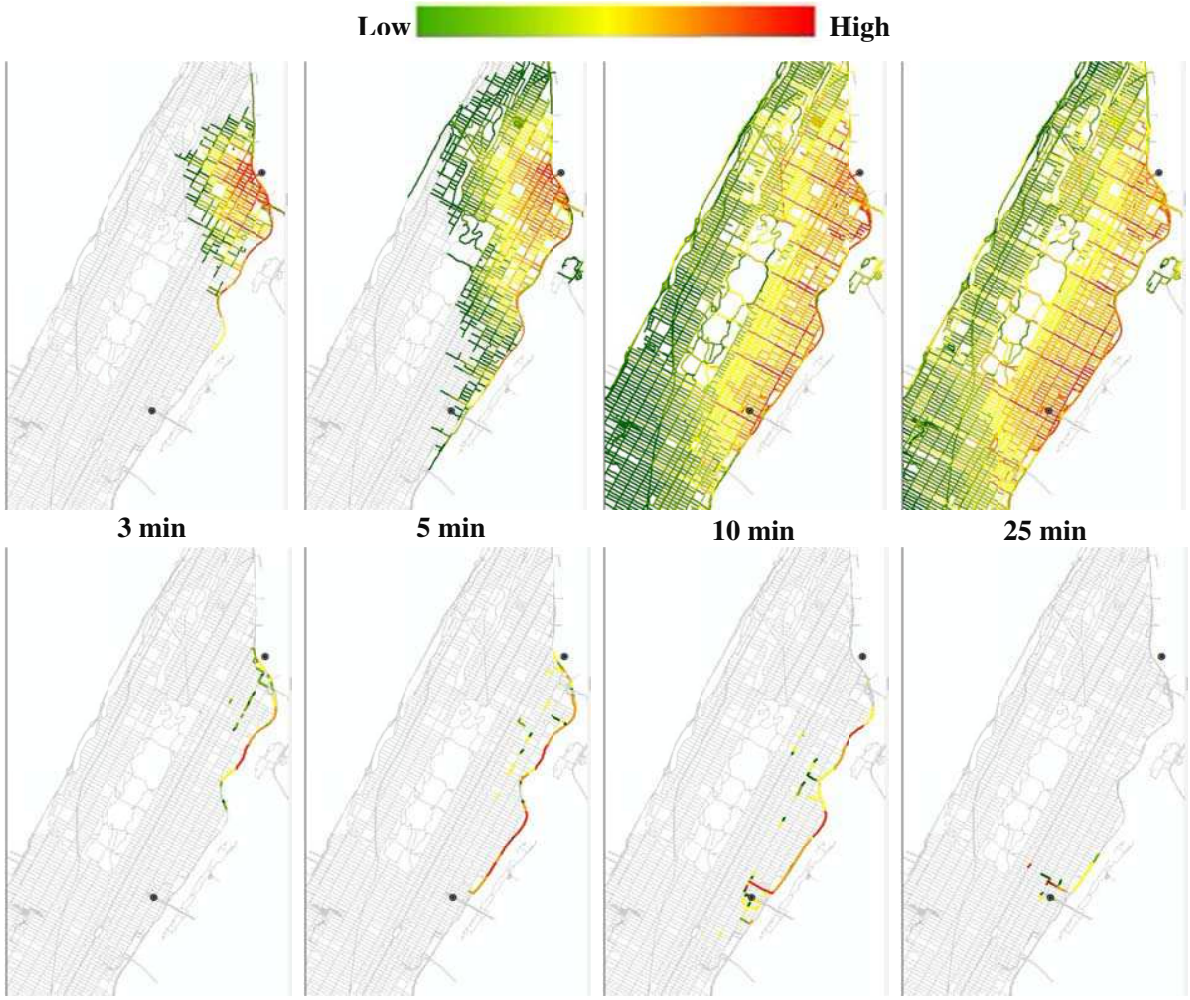



Figure 3. Comparison of simulated and empirical visit probabilities for the southern Manhattan scenario.

## Acknowledgements

This paper is based on research supported by National Science Foundation grant no. BCS-1224102 "Measuring the Environmental Costs of Space-time Prisms in Sustainable Transportation Planning" and the Strategic Highway Research Program SHRP-2 project L04 "Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools."

## References

Kuijpers B Othman W, 2009, Modeling uncertainty of moving objects on road networks via space-time prisms. *International Journal of Geographical Information Science*, 23: 1095-1117.

Miller H J, 1991, Measuring accessibility using space-time prism concepts within geographic information systems. *International Journal of Geographical Information Systems*, 5: 287-301.

Song Y Miller H J, 2014, Simulating visit probability distributions within planar space-time prisms. *International Journal of Geographical Information Science*, 28: 104-125.

Winter S Yin Z C, 2010a, The elements of probabilistic time geography. *Geoinformatica*, 15: 417-434.

Winter S Yin, Z C, 2010b, Directed movements in probabilistic time geography. *International Journal of Geographical Information Science*, 24: 1349-1365.

# Exploring the geo-temporal patterns of the Twitter messages

Muhammad Adnan, Paul Longley
University College London, Department of Geography, Gower Street, London, WC1E 6BT.
Email: m.adnan@ucl.ac.uk ; plongley@geog.ucl.ac.uk

## 1. Introduction

Microblogging services such as Twitter allow users to share information via short messages. Different microblogging services are used not only for communicating with friends, family, and colleagues, but also for real-time news feeds and content sharing about venues (Pennacchiotti and Popescu, 2011). According to recent figures, the Twitter service has more than 200 million active users around the world (Twitter, 2012a). Its major user base is in European countries: in the context of the present paper, usage in the city of London, New York and Paris is the 3rd, 5th, and 7th highest in the world (Bennett, 2012). Twitter users generate a huge quantity of data every day, and our motivation here is to explore the geo-temporal patterns which exist in the text messages themselves. This paper presents an analysis of a large dataset of Twitter messages by the identification of a range of interesting words. Words were assigned to different categories and an initial exploration of the spatial and temporal pattern of the categories is presented.

Analysis of the social media messages is a promising research area. Some related work includes: the use of social media messages to classify areas into homogeneous groups (Birkin et el, 2013), the analysis of the personal information included in the tweet messages (Humphreys et el, 2013), historicizing Twitter within a longer historical framework of diaries (Humphreys et el, 2014), the content analysis of Tobacco-related Twitter posts (Myslín et el, 2012), and a forecasting model to predict the spread of a news (Naveed et el, 2011).

This paper is comprised of 5 sections. Section 2 of this paper describes the data used in the analysis. Data processing is described in the section 3, while section 4 and 5 present the results and conclusion.

## 2. Data

The Twitter Streaming API (Twitter, 2012b) can be used to download a 1% sample of the geotagged tweets. For this paper, the Twitter Streaming API was used to download geo-tagged Tweets for the Greater London during July to December, 2013. The fields downloaded from the API included the user name, latitude and longitude from which the Tweet was sent, time and tweet message content. A total of 4.6 million (4,633,139) geo-tagged Tweets were downloaded. These tweets were sent by a total of 272,248 unique users.

Few users sent more tweets than others. 2,000 or more tweets were sent by the top 45 users. Following figure (1) shows the number of tweets by individual users.

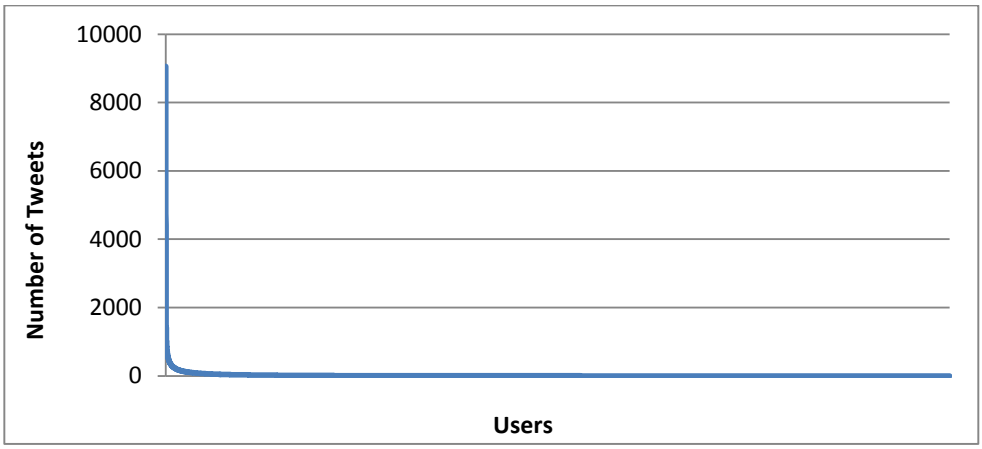



Figure 1: Number of tweets by individual users

## 3. Data processing

In the first step, 4.6 million tweets messages were divided into a series of 'words' i.e. a group of characters separated by a full-stop, comma, semi-colon, colon, apostrophe, or double quotes. This resulted in a dataset of 35,028,273 words. For the investigation of the spatial patterns of individual words, all the words were aggregated to 633 wards in the Greater London. For each word (y), an Index of Dissimilarity (Birkin et el, 2013) was calculated across the 633 wards. The Index of Dissimilarity is defined in the following equation.

$$\theta\,(x,z) = 0.5 \times \sum_{x} \left| \frac{X_x^y}{X_*^y} - \frac{X_x^*}{X_*^*} \right|$$

Where $x = (1,\ldots..,633)$ wards in the Greater London and an asterisk (*) denotes summation across a missing index. The resulting Index of Dissimilarity value for each word (y) is a standardized value between 0 and 1. Where 0 indicates a uniform distribution and a 1 indicates a spatial concentration.

Index of Dissimilarity was calculated for each of the 35,028,273 words in the dataset. In the second step, in order to select the words which are spatially concentrated, the words having Index of Dissimilarity less than 0.5 were deleted from the database. This resulted in 122 remaining words which are listed in the following table (1). The table also assigns each word to one of the 8 distinct categories.

Table 1: 122 spatial concentrated words

| Categories | Words |
|---|---|
| **Travel** | LHR, PANCRAS, PADDINGTON, HEATHROW, RAILWAY, UNDERGROUND, FLIGHT, STATION, @HEATHROWAIRPORT, AIRPORT, TERMINAL, TUBE |
| **Sports** | #THFC, FULHAN, #ARSENAL, #LFC, #AFC, #ASHES, #CFC, @ARSENAL, CHELSEA, SPURS, FOOTBALL, #MUFC |
| **Places in London** | HOUSNLOW, MARYLEBONE, MIDDLESEX, BROMLEY, GREENWICH, ISLINGTON, SHOREDITCH, OXFORD, PICCADILLY, WHARF, KINGSTON, SHARD, HACKNEY, BRIXTON, BRICK, MARKET, KENSIGNTON, LEICESTER, KNIGHTSBRIDGE, CROYDON, HAMMERSMITH, CIRCUS, TOTTENHAM, WATERLOO, NOTTING, COVENT, REGENT, ARENA, WESTFIELD, ROMFORD, CAMDEN, RICHMOND, CLAPHAM, STRATFORD |
| **Tourism** | MUSEUM, TOWER, GALLERY, BRIDGE, PALACE, ROYAL, HOTEL, COURT, TRAFALGAR, HYDE, WESTMINSTER, ALBERT, BUCKINGHAM |
| **Food & Drink** | @STARBUCKSUK, STARBUCKS, COCKTAILS, BAR, COSTA, PUB, DRINK, COFFEE, JUICE, CAFE, MCDONALDS, COOKING, RESTAURANT |
| **Leisure** | LOUNGE, STUDIOS, THEATRE, PARK, EVENT, CINEMA, XFACTOR, KITCHEN, HOLIDAY, XBOX, HANGING, GARDEN, SHOPPING, MUSIC |

| | |
|---|---|
| **Emotions** | ENJOYED, #EXCITED, OMG, MISSING, SURPRISED, DISGUSTING, EMBARRASING, ANNOYING, GAY, MADNESS, WTF, FANTASTIC, SHOCKING, RIDICULOUS, BORED, AWFUL, HAPPINESS, PLZ |
| **Other** | GOODNIGHT, DUDE, DAD, DADDY, BOYS, FAMILY, FRIEND |

## 4. Results and Discussion

Following figure (2) shows an example of the spatial concentration of the words. This figures shows two maps of the individual tweets where 'TRAFALGAR' (map on the left) and 'LHR' (map on the right) were mentioned in the tweet messages. The Index of Dissimilarity value for both the words was 0.833 and 0.96 respectively, indicating a spatial concentration of the tweets.

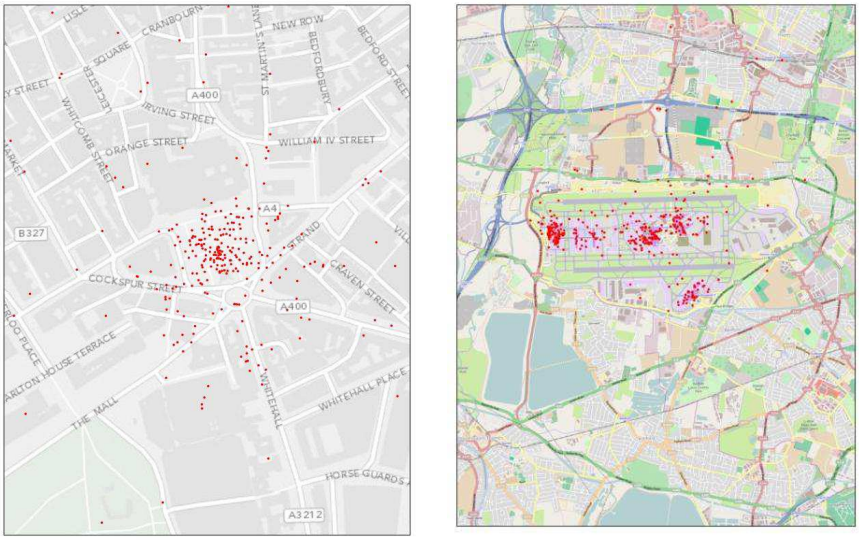
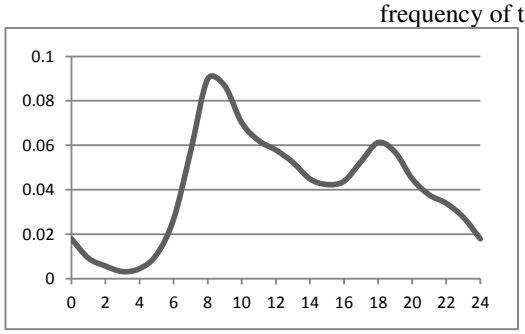


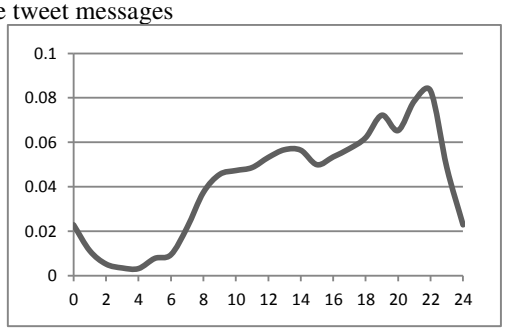

Figure 2: Tweets around the area of Trafalgar Square (left) and London Heathrow Airport (right)

The following table (2) shows the temporal graphs of the 4 word categories listed in section 3. The temporal graphs show the distinct temporal patterns of these categories. Words of the 'Travel', 'Sports', and 'Leisure' categories have the most distinct patterns. There is high number of tweets mentioning 'Travel' category words during the morning and evening rush hours. More tweets of the 'Sports' and 'Leisure' category words are sent during the night time. There are also more tweet mentions of the tourist places after 3pm during the day.

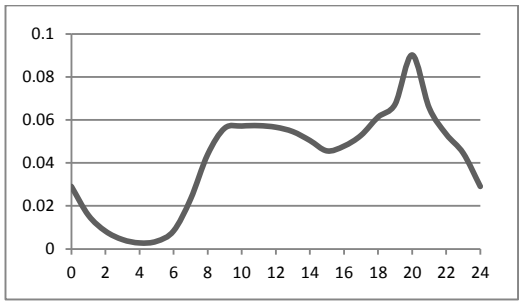
Table 2: Temporal graphs of the word categories. X-axis represents the hours of the day and Y-axis represents the frequency of the tweet messages



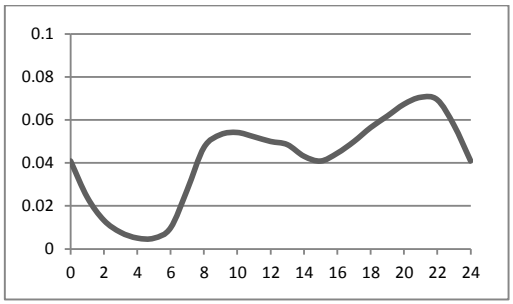

Travel




Sports




Leisure




Tourism

## 5. Conclusion and future work

This paper has presented a preliminary analysis of the Twitter messages to explore the inherent spatial and temporal patterns of activity. A large dataset of Twitter messages was analyzed and decomposed into 35,028,273 words. For each word, the Index of Dissimilarity was calculated to identity interesting words having spatial concentrations. This resulted in a total of 122 words which were assigned to 8 distinct categories. The paper has also presented an initial exploration of the spatial and temporal pattern of the word categories.

This is a very promising research area and there are a number of ways in which this work could be improved in the future. A possible improvement is to perform a fine scale temporal activity pattern analysis on the dataset to identity the areas of distinct attributes and behaviors e.g. the areas of leisure activities vs. work place areas.

## Acknowledgements

This work was completed as part of the EPSRC research Grant "The Uncertainty of Identity: Linking Spatiotemporal Information in the Real and Virtual Worlds" (EP/J005266/1).

## References

Bennet, S. 2012. Revealed: The Top 20 Countries and Cities of Twitter [STATS]. Retrieved 31st December, 2012, from http://www.mediabistro.com/alltwitter/twitter-top-countries_b26726.

Birkin, M., Harland, K., Malleson, N. (2013). The classification of space-time behavior patterns in a British city from crowd-sourced data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7974, pp.179-192.

Humphreys, L., Gill, Phillipa., Krishnamurthy, B. 2013. Historicizing New Media: A Content Analysis of Twitter. Journal of Communication, 63, 413-431.

Humphreys, L., Gill, Phillipa., Krishnamurthy, B. 2014. Twitter: a content analysis of personal information. Information, Communication & Society. 17 (7).

Myslín, M., Zhu, Shu-Hong., Conway, Michael. 2012. Content Analysis of Tobacco-related Twitter Posts. In the proceedings of the 2012 International Society for Disease Surveillance Conference.

Pennacchiotti, M. and Popescu, A. 2011. A Machine Learning Approach to Twitter User Classification. In Proceedings of the Fifth International AAAI conference on Weblogs and Social Media.

Naveed, N., Gottron, T., Kunegis, Jérôme., Alhadi, Arifah Che. 2011. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In the proceedings of the WebSci'11. Koblenz, Germany. June 14-17, 2011.
Twitter. 2012a. What is Twitter ?. Retrieved 31st December, 2012, from https://business.twitter.com/basics/what-is-twitter/.

Twitter. 2012b. The Streaming APIs ?. Retrieved 22nd January, 2012, from https://dev.twitter.com/docs/streaming-apis.

# Building a CityGML Infrastructure for Energy Related Simulations

A. N. Alexandru Nichersu, A. S. Alexander Simons

[1]EIFER European Institute for Energy Research, Emmy-Noether-Strasse 11, Karlsruhe, Germany
Email: (alexandru.nichersu, alexander.simons)@eifer.uni-karlsruhe.de

## 1. Introduction

Energy plays a vital role in the world's economy. Heat, light and power allows cities and factories to provide jobs, goods and homes (Voser, 2012). To be able to provide the energy we need to understand the demand, locate and simulate and analyse its variation in size and time. Simulations are required to provide decision support on matters such as improving the efficiency of energy use or to forecast energy demand in the future. According to (Bahu, 2013) energy system models capable of describing future energy systems require a spatial representation in order to reflect the local context and the boundary conditions.

CityGML is, in the words of its creators (Kolbe, 2009), an open data model and XML-based format for the representation and exchange of virtual 3D City models. It complements existing standards in the Geomatics field and has been seldom used in the energy field. As it is a fairly new standard new methodologies and infrastructure have to be developed to make best use of it and this was the purpose of this work. We decided to choose CityGML because its development is aimed for reaching a common definition of basic entities, attributes and relations of 3D city models. This allows the reuse of the same data in different application fields (OGC, 2012).

## 2. Infrastructure developments

The objective of the project was to create an open and flexible 3D data infrastructure to connect to different simulation tools and integrate calculation methods. The individual components are described in the following section.

### 2.1 Converting to CityGML

Energy simulations are run in different environments depending on the language the model was modelled and simulated in. Our objective was to create an environment that can be connected to most simulation environments so that the geo-localized data can be stored, accessed, transformed and if needed, visualized. This is why we decided to use PostgreSQL databases.

To add geo-localized data to the PostgreSQL databases we used the extension called PostGIS. PostGIS allows the DB to store geometry and to perform spatial calculations on the data.

Most of the geo-localized data in the field of energy modelling is currently available as shapefiles or SketchUp files. Therefore a conversion to the CityGML standard is needed. Our aim for the conversion was a LoD 2 (Level of Detail) model which only describes the outside shell of the building with deflections or the roof shape. To be able to complete the standard conversion we have used two different solutions.

The first solution was developed using FME Workbench. This requires a DTM (digital terrain model) and the footprint of buildings (with height present in the shapefiles as an attribute).

The second solution is the CityGML-Editor (SketchUp CityGML-Plugin) of the Westfälische Hochschule that converts SketchUp format directly to CityGML (CityGML-Toolchain, 2013). This only requires the building to be built in SketchUp to make the conversion and the surfaces defining the building to be created in three separate layers: ground surface, wall surface, roof surface.

## 2.2 Uploading CityGML to the PostgreSQL database

The newly generated CityGML files have to be uploaded to the database for further analysis. To do this we used the 3D City DB importer/exporter tool. This tool is a free 3D geo-database to store, represent, and manage virtual 3D city models on top of a standard spatial relational database (Stadler et al. 2009).

The tool first creates a relational geo-database schema which is based on a simplification of the CityGML's data model that still maintains the key features of CityGML. The tool also provides the user with XML validation of CityGML documents which has proven to be an important asset especially after the conversions from other standards. During the import of CityGML files to the DB resulted from different data sources the tool validates the XML code in the file and provides error information if any is found.

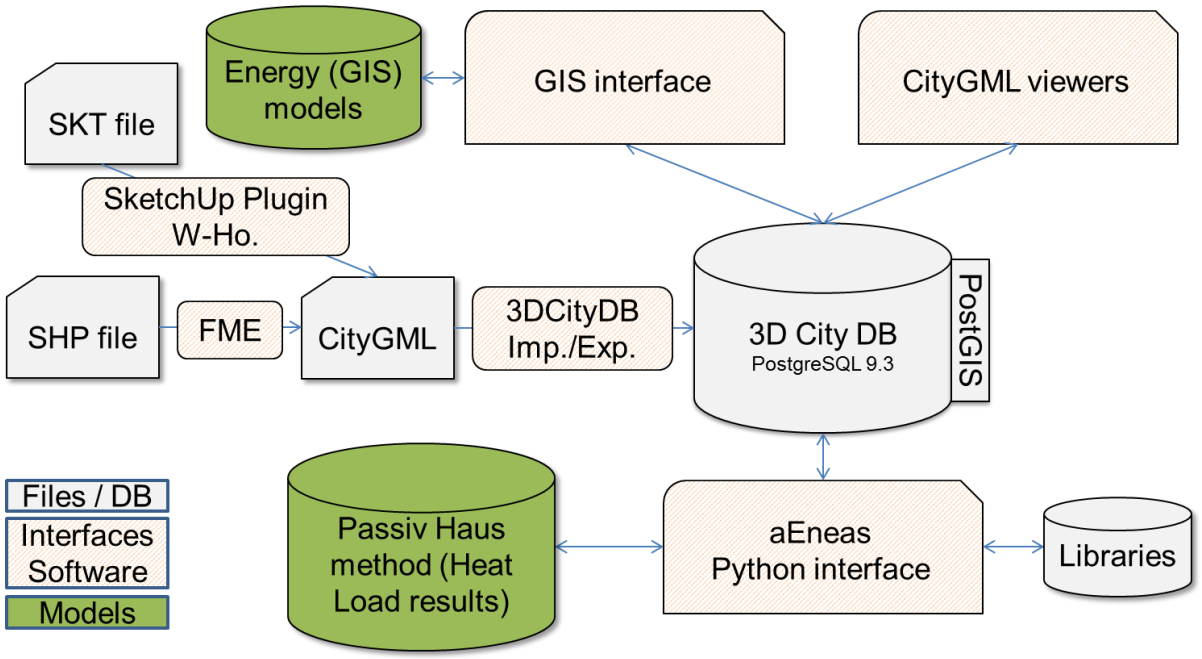



Figure 1. Infrastructure and workflow

After the data is imported in the database we integrated the energy related models to our new system. As a system component it can now house energy model parameters within the CityGML format.

## 2.3 Connecting energy related models to the DBMS

Figure 1 shows the energy related models could be separated in two groups, GIS environment models and programming environment models.

Previous models integrated in GIS environments would connect to the geospatial data directly from the GIS system. Within our new infrastructure it was possible via an ODBC (Open Database Connectivity) link to our DBMS (database management systems) to connect the geo-information directly to the GIS system. This in turn allowed previously developed models to change from shapefiles to CityGML without much further work.

The programming environment models allow connectivity to the DBMS also via ODBC - Python, MATLAB or JDBC (Java Database Connectivity) - Python, and Anylogic. As a proof of concept we used a model called the EBM (Energy Balance Model) which is developed in Python.

## 3 Energy Balance Model

As an application case an Energy Balance Model was used to test the applicability of the developed framework.

The Energy Balance Model is based on the PassivHaus method developed by the Passiv Haus Institute (Feist 2007). From the LoD2 building model in the database we determine the volume and the different areas of roof, wall and ground surface. The previously mentioned surfaces together form the building envelope, which can be used for calculating energy gains and losses. The various surfaces are multiplied with U-values for their corresponding types. The U-values are taken from the Institute for Housing and Environment GmbH (ger.: Institut Wohnen und Umwelt, IWU) documentation (IWU, 2005). This document describes the energy saving potential by thermal protection measures for buildings in Germany. By using these guidelines, we classified the building stock by the age and size, which are important indicators for U-Values. Using the values calculated in the previous step, the transmission heat loss and ventilation heat loss are calculated. After this initial step, the temperature zone factor and air capacity coefficient are taken into account. The calculation of the heat demand uses the following formula and the results are given in W/K:

$$Demand_{hl} = \sum(A_i * U_i * f_T) + (V_A * n_A * c_{Air}) \quad (1)$$

Where:

| | | |
|---|---|---|
| $A_i$: | Component surface area | m |
| $U_i$: | U-value of surface component | w/m$^2$ |
| $f_T$: | Factor for temperature zone | |
| $V_A$: | Air volume | m$^3$ |
| $n_A$: | Energy-efficient air exchange in the heating load case | |
| $c_{Air}$: | Heat capacity of air | J/Kg |

Afterwards a Heating Degree Day (HDD) is calculated in Kh. Heating Degree Day is a value determined by the difference between the inside temperature of a building $T_i$ and the daily outside temperature $T_o$. This is done for every day on which the mean outside temperature is lower than the heating limit. In Germany the used limit temperature is often given with 15°. The summation of all days of a year delivers a measurement designed to determine the required energy demand for heating a building.

$$HDD_a = \sum_1^{12}[(H_D * D_M) * (T_i - T_o)] \quad (2)$$

Where:

| | | |
|---|---|---|
| $H_D$: | Heating hours a day in 24 | h |
| $D_M$: | Number of days per month | |
| $T_i$: | Inside temperature of building | ºC |
| $T_o$: | Average outside temperature per month | ºC |

For the energy demand per building in kWh we use the following formula:

$$Energy\ demand_{bui} = Demand_{hl} * HDD_a * \frac{1}{1000} \quad (3)$$

The method applied in our current infrastructure allowed the calculation of the energy demand for a neighbourhood of the city of Karlsruhe (365 LoD2 buildings) in 25.1 seconds. The machine specifications are as follows:

System: Ubuntu release 12.04(precise) 64-bit
Kernel Linux 3.11.0-15-generic
GNOME 3.4.2

Hardware: Memory: 7.8 GB
Processor: Intel® Xeon® CPU E5-2690 0 @ 2.90GHz x 16

## 4 Conclusion

The study has shown that the infrastructure developed proved to be flexible and adaptive to the requirements set by the energy models. As such we were able to successfully implement an energy balance model and get an assessment of the results.
The coupling of the CityGML standard, structure of 3DCityDB and the energy models gives us the freedom to model at different spatial scales, include information that was previously unavailable directly to the models, such as DTM, building texture or using spatial queries and make easy use of all the advantages that geo-localized information can bring to our models.

## Acknowledgements

The authors of this study would like to thank their colleagues at EIFER – the European Institute For Energy Research and KIT–IPF – Karlsruhe Institute of Technology – Institut für Photogrammetrie und Fernerkundung for their continued support and assistance. The methodological basis presented here was developed in the context of a research project called “Entwicklung und Validierung interferometrischer und radargrammetrischer Bildanalysemethoden zur automatisierten Extraktion und Charakterisierung von 3D-Gebäudestrukturen für energie- und krisenrelevante Geoinformation” which has been running from 2012 to 2014 in the two institutions of the KIT.

## References

Peter Voser, Energy: The Oxygen of the Economy, Industry Agenda, World Economic Forum, 2012
Thomas Kolbe, Representing and Exchanging 3D City Models with CityGML, 2009
Wolfgang Feist et al., PHPP Passivhaus Projektierungs-Paket 2007
J.-M. Bahu*, A. Koch, E. Kremers, S.M. Murshed, Towards a 3d spatial urban energy modelling approach, 2013, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume II-2/W1, ISPRS 8th 3DGeoInfo Conference & WG II/2 Workshop, 27 – 29 November 2013, Istanbul, Turkey
Alexandra Stadler, Claus Nagel, Gerhard König, Thomas H. Kolbe, Making interoperability persistent: A 3D geo database based on CityGML, 2009
Institut Wohnen und Umwelt GmbH, Deutsche Gebäudetypologie Systematik und Datensätze, 2005
Open Geospatial Consortium, OGC City Geography Markup Language (CityGML) Encoding Standard, 2012
CityGML-Toolchain Editor 1.8, 2013

# Determining Hierarchy of Landmarks in Spatial Descriptions

Vanessa Joy A. Anacta, Angela Schwering, Rui Li

Institute for Geoinformatics, University of Muenster
Heisenbergstrasse 2, 48149, Muenster, Germany
Email: {v.anacta; schwering; rui.li}@uni-muenster.de

## 1. Introduction

Communicating wayfinding instructions involves selective inclusion of spatial elements that may guide a person to reach the destination. These spatial elements could provide someone an idea of the place layout depending on how instructions are structured. Allen (1997) discussed methods of how a person is able to communicate route instructions well in ways that are easily comprehensible that could also be applied to navigation systems. This means that route instructions should be appropriate to a certain environment as well as the type of people. The quality of route instructions is important for effective wayfinding, but having lengthy or brief instructions does not translate into either good or bad verbal route instructions (Lovelace et al, 1999). Moreover, Weissensteiner and Winter (2004) investigated the importance of narratives in providing wayfinding instructions as it engages the person to the environment thereby, creating a picture of the unfamiliar area and its surroundings. Without landmarks, it will be hard for people to find their way especially those who do not prefer following absolute directions. Raubal and Winter (2002) addressed the importance of enriching wayfinding instructions with local landmarks by providing measures to identify the saliency of a specific feature. Richter and Klippel (2005) highlighted that the structure of the environment plays a major role on how wayfinding instructions should be written. In this study, we investigate the types of landmark information participants include in verbal route descriptions and sketch maps. We focus our analysis on the composition of landmarks in the spatial descriptions whether participants are confined in giving only local landmarks in the route instructions.

### 1.1 Types of landmarks

In this paper, we classify landmarks into local and global landmarks. Local landmarks refer to landmarks along the route (LLAR) or landmarks at decision points (LLDP) with turning action. Local landmarks are mostly used in today's navigation instructions guiding people in a new environment by following turn by turn directions. But, people may tend to include global landmarks that are also helpful reference objects in wayfinding. Global landmarks (GL) are identified as either point or regional features situated off-the route. Point-like features refer to specific buildings while regional features are landmarks with an area extent (e.g. lake, mountain, city center). These may not necessarily be visible landmarks which are located along the route but they could also be point or regional features that are distant but could be useful information for orientation. Distant landmarks which are less exploited in verbal instructions provide someone global orientation (Couclelis, 1996; Winter et al, 2008) which might help one capture a survey knowledge of an unfamiliar environment.

### 1.2 Hierarchy in spatial descriptions

Hierarchy of spatial objects is evident on how landmarks and paths are clustered based on its functionality. This has been investigated in the development of the anchorpoint theory (Golledge, 1997). As what the author emphasized, anchorpoints do not only refer to well-known and mostly used place in the environment. Taylor and Tversky (1992) found out that there was a correlation between the order of elements drawn and the order on how it was mentioned in the spatial descriptions. It occurred in their study that there is hierarchical structure in people's sketch maps at different environmental scales.

Extensive research on global landmarks is limited. Steck and Mallot (2000) developed a virtual environment and looked at how people refer to local and global landmarks in the navigation task. In this particular task, the authors defined global landmark as a reference frame which does not change if the participant move even at short distance. Examples of the global landmarks used were mountain, city skyline and TV tower. Local landmarks, on the other hand, refer to visible objects near the route and seen from a small distance. It resulted that both local and global landmarks were used for wayfinding decision tasks. Winter et al (2008) developed an approach showing hierarchical communication of space through partitioning of landmarks. Landmarks have been considered either a point in the route or a component of a region. In their study, a wayfinding instruction was developed such that the person is first directed to a prominent feature and from there instructions to the real destination were provided.

## 2. Results

We asked 17 university students in Muenster, Germany to provide wayfinding instructions to someone unfamiliar of the city through text and sketch map. From the initial result, all

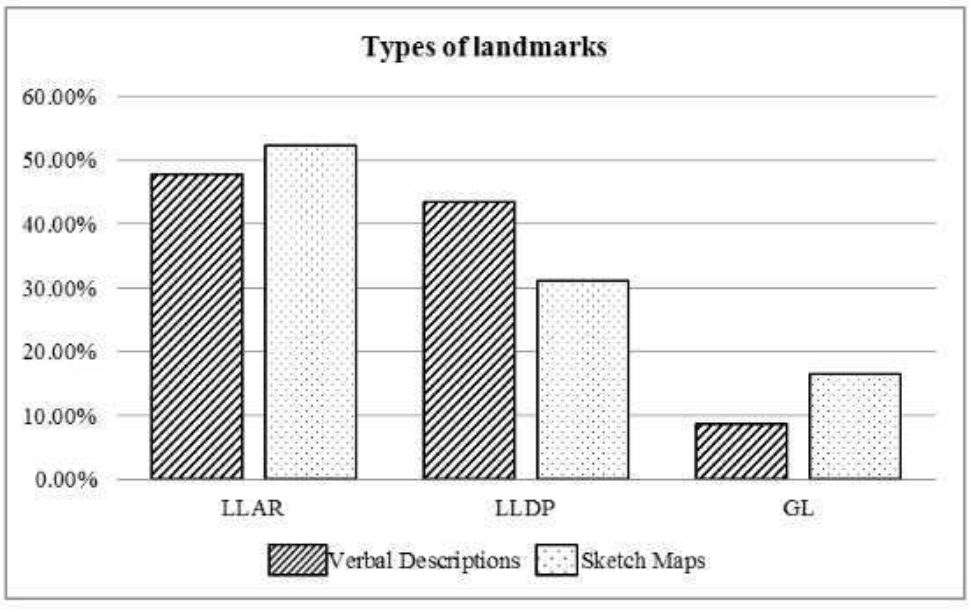



Figure 1: Frequency of landmarks in verbal descriptions and sketch maps.

participants included both types of local landmarks along the route and at decision points. Majority of the participants, 70.59% and 76.47% have included global landmarks in their verbal descriptions and sketch maps, respectively. Figure 1 shows the frequency of the types

of landmarks included in both spatial descriptions with more local landmarks along the route and followed by local landmarks at decision points and finally, global landmarks.

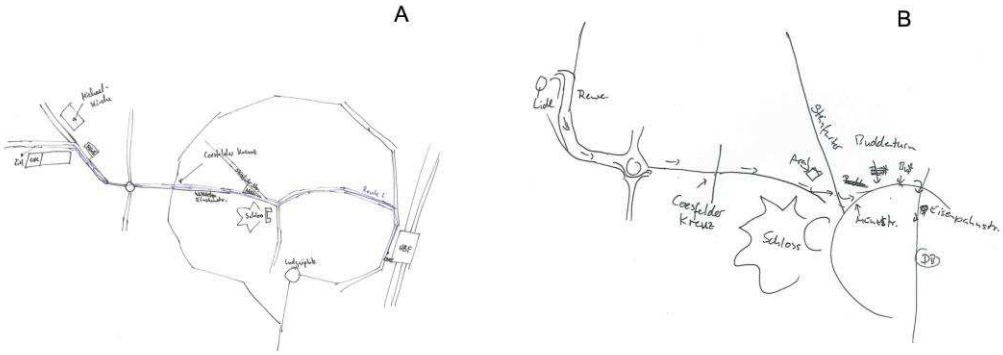



Figure 2: Example of sketch maps with global and local landmarks.

Prominent landmarks are oftentimes used as reference objects in providing instructions. In the case of Muenster, the castle (Schloss) and the Promenade were the prominent landmarks frequently used as reference objects in giving wayfinding instructions. The Promenade defines a boundary for people to refer to the location of the city center which is basically described as the area inside it (see Fig. 2).

Given that the participants combined local landmarks and global landmarks in their wayfinding instruction, we took this into account in our preliminary analysis of identifying presence of hierarchy in the spatial descriptions and incurred the following observations:

a) Participants provide a global orientation by giving a distant region which does not necessarily have to be along the route but providing an idea of the direction of travel. One example states: *"*From there, you drive to the direction of the castle. But you turn left before *the castle".*

b) Participants summarize the route and describing afterwards what other landmarks and streets to see along the route in sequence.

c) With regard to sketch maps, some maps were structured showing a 'landmark within a landmark' concept wherein a landmark point feature is a component of a regional landmark as shown in Figure 2a.

## 3. Conclusion and Future Work

The result of the preliminary analysis suggests that both local and global landmarks have been used in giving route descriptions. It showed that participants include more landmarks along the route where there is no turning action. We observed also an inclusion of off route distant landmarks in both spatial descriptions. Global landmarks could be considered important in wayfinding but its function and potential use for during wayfinding is not extensively studied.

We are currently investigating how to systematically analyze the hierarchical structure of objects in spatial descriptions. The number of participants that have included global landmarks in the sketch maps is an indicator suggesting that these are also important elements in giving wayfinding instructions. This aspect is not extensively investigated which is why we consider the importance of combining global landmarks and local landmarks in

wayfinding instructions. We find this approach valuable in developing more meaningful instructions that may instill spatial layout learning and not only focusing on procedural steps.

Furthermore, we intend to explore the role of global landmarks in more details. In our present study, we grouped distant landmarks (either regional or point-like) that may serve the role of maintaining orientation in category of global landmark. It will be worthwhile for us to further explore the roles of different distant landmarks due to their shape, location, or distance.

## Acknowledgements

We acknowledge the support of the DAAD and the DFG-funded projects – SketchMapia and WayTO.

## References

Allen, G.,1997, From knowledge to words to wayfinding: Issues in the production and comprehension of route directions. In Hirtle, S., Frank, A., eds. : Spatial Information Theory A Theoretical Basis for GIS 1329. Springer Berlin Heidelberg, 363-372.

Couclelis, H., 1996, The Construction of Cognitive Maps. Springer Netherlands,133-153.

Golledge, R., Stimson, R. J., 1997, Spatial Behavior: A Geographic Perspective. The Guilford Press, New York.

Lovelace, K., Hegarty, M., Montello, D, 1999, Elements of Good Route Directions in Familiar and Unfamiliar Environments. In : COSIT '99 Proceedings of the International Conference on Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science, pp.65-82.

Raubal, M., Winter, S., 2002, Enriching Wayfinding Instructions with Local Landmarks. In Egenhofer, M., Mark, D., eds. : GIScience '02 Proceedings of the Second International Conference on Geographic Information Science, pp.243-259.

Richter, K.-F., Klippel, A.,2005, A Model for Context-Specific Route Directions. In Christian Freksa, M., ed. : Proceedings of the 4th international conference on Spatial Cognition: reasoning, Action, Interaction, pp.58-78.

Steck, S., Mallot, H.,2000, The role of global and local landmarks in virtual environment navigation. Presence: Teleoperators and Virtual Environments Vol. 9 Issue 1, 69-83.

Taylor, H., Tversky, B., 1992, Descriptions and depictions of environments. Memory and Cognition Vol. 20, Issue 5, 483-496.

Weissensteiner, E., Winter, S.2004, Landmarks in the Communication of Route Directions. In : Third International Conference, GIScience 2004, Lecture Notes in Computer Science, vol. 3234, pp.313-326.

Winter, S., Tomko, M., Elias, B., Sester, M.,2008, Landmark hierarchies in context. Environment and Planning B: Planning and Design 35 (3), 381-398.

# Is Thematic Uncertainty Beneficial?

# Decision-Making in Air Pollution Health Alerts—Qualitative Research

R. Bacova

Department of Geography, Faculty of Science, Masaryk University, Kotlarska 267/2, 611 37, Brno
Email: rada.ba@mail.muni.cz

## 1. Introduction

Several times a day, each of us must make a decision in a given situation. An example is the decision process that leads to geographical consequences. Our knowledge is a prerequisite for making good, effective decisions, as well as for creating interoperable and accurate geographical data and maps. The concepts of vagueness and accuracy in geography can be found in many terms that are more or less semantically related, such as data quality, errors and uncertainty. At the most general level, there is the uncertainty of visualized and communicated information, considered one of the critical factors of cartographic production (Kubíček and Šašinka 2011).

Maps are the main result of cartography and can be regarded as finite presentations of an infinite reality, i.e. the model of a large reality. Inherently, therefore, they must inevitably simplify the actual situation in order to present it to the end users. The effectiveness of such decisions is affected not only by the quality of the data and information encoded in the map but by the cartographic communication paradigm (Koláčný 1969, Morita 2004), as well as the knowledge and experience of the map reader. Due to technical reasons, it is impossible to record all known facts on maps; thus, the cartographer must limit them to the most important ones. Similarly, for the decision-making process, since it is impossible to take into account all the knowledge we have about a context, it is appropriate to address some characteristic of data accuracy and analysis accuracy, such as uncertainty.

The aim of this paper is to present an empirical research design intended to determine the extent to which the visualization of uncertainty influences decisions in real situations. The factual context is the announcement of outgoing restrictions due to excess $PM_{10}$ concentrations and smog formation in municipalities in the Czech Republic.

## 2. Theoretical Background

### 2.1 Uncertainty in geographical analysis

Numerous research and case studies have been published on the extension of the concept of uncertainty in geography in the last few years. They represent uncertainty as an appropriate element of cartographic visualization to improve the quality of information and capture its spatial variability. Unfortunately, several different views exist regarding the actual meaning of uncertainty (MacEachren 1995, Pang 1997, Zhang and Goodchild 2002, Hall 2003, Longley 2005). The author of this text understands uncertainty (taxonomy is in Figure 1) as the complementary characteristics of gross errors, which together form the imprecision of geographic information. Uncertainty can be further divided into inherent and epistemic types, while cartographic production reflects only inherent uncertainty, which can be detected and

quantified (Hall 2003). Moreover, Pang (1997) added the positional, attribute (thematic) and temporal conditionally classifications of uncertainty. This paper takes into account spatially related thematic uncertainty, which is the most relevant type of uncertainty in the investigated phenomenon. The aggregate representation of spatial and thematic uncertainty is necessary due to the lack of justification for the identification of the separate effects of each type in their geographical scope.

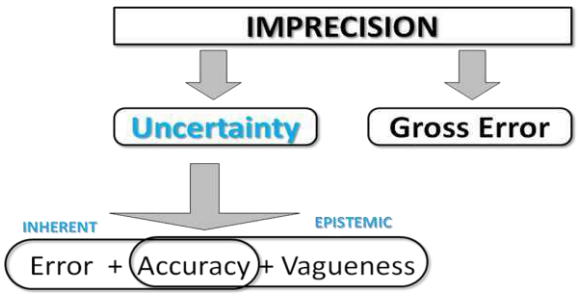



Figure 1. Taxonomy of uncertainty.
(Based on MacEachren 1995, Zhang and Goodchild 2002, Hall 2003; adapted by author)

### 2.2 Air pollution health alert

The acronym $PM_{10}$ refers to the particulate matter described in Regulations (EC) No. 166/2006 of the European Parliament and of the Council of 18 January 2006 concerning the establishment of a European pollutant release and transfer register and amending Council Directives 91/689/EEC and 96/61/EC. The Directive also specifies the obligation of EU Member States to determine the degree of concentration of aerosol particles with the highest precision possible, as these are products of human activity which negatively affect human health. The concentration of airborne dust is measured on two types of land-based weather stations which are divided according to their background, i.e. urban and rural. Due to the accumulation of dust in the lungs causing the deterioration of the already sick population, small children and the elderly, there is a set concentration limit at which the measures are made public.

The population is warned about adverse air quality and is given the following recommendations: limit ventilation to five minutes; reduce time spent outdoors; the elderly, the sick and children should not go outdoors at all; do not use motor vehicles; use air cleaners and reduce smoking when exceeding the concentration limit and defined forecast's durability. The Czech Hydrometeorological Institute (CHMI) announces the aforementioned information which is put into practice by the administrative authority.

## 3. Experimental Design

The research includes a series of case studies measuring the ability of users to work with different methods of cartographic visualization. A basic study of the cognitive processes in cartography adds the issue of spatial decision. The discussed design focuses on identifying the effectiveness and efficiency of the real decision-making in a smog situation. A similar study can be found, for example, in the work of Senaratne et al. (2012). The primary objective of this study is not to identify user preferences in graphical variables for cartographic visualization but to uncover the effects of the inclusion and awareness of uncertainty on the decision-making process, based on specific, real needs.

The research design follows the conventions of mixed methods research and comprises two complementary research methods. The methods are a qualitative pre-test represented by structured interviews with experts, followed by a quantitative test of experts and cartographers supplemented by other qualitative methods including a survey which will be

administered after the completion of the quantitative test. The basic hypothesis is as follows: The explicit display of uncertainty results in clearer, more accurate and more objective decision making, thus limiting subjective judgements arising due to the personal characteristics of the respondents.

The basic task is for the respondents to define the administrative areas in which a curfew would be declared in response to the critical threshold concentration of $PM_{10}$, i.e. 50 $g*m^{-3}$, being exceeded. The qualitative pre-test (interviews) is complemented by consultations to explain possible ambiguity and the concept of uncertainty. The respondents chosen for the interviews were selected due to their high competence in air pollution health alerts and meteorological map interpretation. They are all employees of CHMI with long work experience, and are also academic staff in the field of meteorology and air pollution. They were asked to participate in a brief discussion about the types of maps used in the test design.

The main quantitative empirical test will be administered in an online environment, MuTeP, and is divided into three randomly rendered sections to avoid the order of the tasks influencing the results. The first section contains the basic work of real thematic mapping that is commonly used as a result of the geographical analysis of ground measurements of $PM_{10}$ in the Czech Republic (an example is shown in Figure 2) with a well-established multi-hue colour scheme. This map is accompanied by a metadata query under the INSPIRE directive containing prescribed information about the data quality.

The next two sections of the quantitative empirical test concern the possible ways to display thematic uncertainty, which were taken from existing studies (Leitner and Buttenfield 2000, MacEachren 2005). The users will be required to compared maps and to combined maps (Figure 2), where the uncertainty is visualized with the most common cartographic method, namely, whitening. The use of both basic methods (compared and combined maps) of displaying the uncertainty will capture individual cognitive personality types, which will significantly affect how the respondents work with the maps. The results of the quantitative test will exhibit the errors and the time required to provide each answer with sufficient statistical power.

After completing this text, some respondents will be given a short survey to complete which will ask them to clarify possible problems and express their level of satisfaction with the different types of visualization in the previous empirical test.

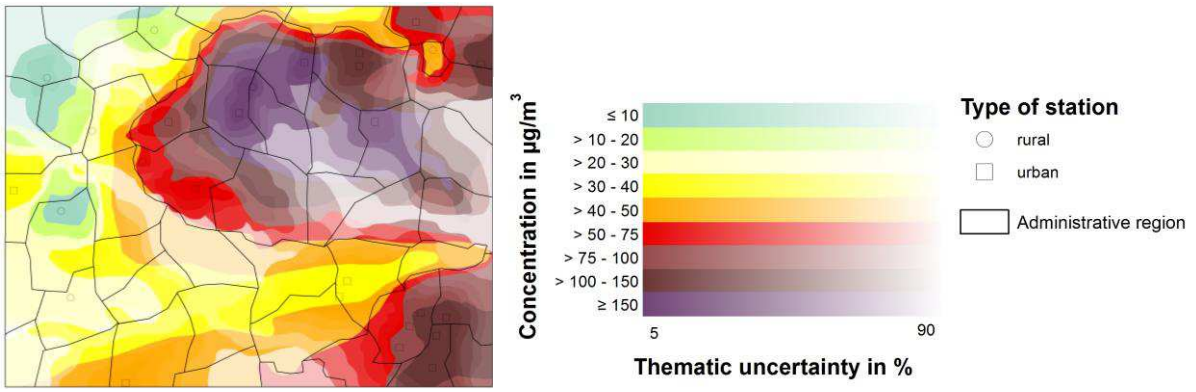



Figure 2. Maps combined (illustration): $PM_{10}$ and thematic uncertainty.

## 4. Expected Results and Future Work

The results obtained from the interviews are expected to provide relevant information regarding the ability to understand and work with uncertainty in the common practice of issuing health alerts. The main goal is not precisely to quantify the suitability of the selected methods of geographic visualization but to determine all possible aspects that influence the

decision making of experts who do not have high cartographic knowledge, but use thematic maps to make decisions. The results will contain a list of recommendations and restrictions to create the resulting thematic uncertainty map, adapted to the conditions of the field of meteorology, which will be used in the next step of the research design. Specifically, the author will consider the respondents' knowledge about uncertainty, whether it can be deemed an appropriate characteristic of maps and data quality, whether it is appropriate for quantifying uncertainty and how the respondents are able to become accustomed to the new method of cartographic visualization.

The qualitative research described in this paper will be completed in the medium term with the outlined empirical test. The main problem which the author has observed in existing studies is the issue of the separation of quantitative and qualitative testing methods and geographic visualization methods used for evaluating this information. The mixed methods research described above is therefore an optimal starting point for overcoming the limitations associated with the methods used in previous studies. In addition, it is necessary to carry out further case studies with real tasks used in practice, to promote the concept of uncertainty in geography. However, at the same time, it is important to note that uncertainty should not be the target in all cases—only when it is beneficial to the decision making. Further, empirical tests should not be oriented only towards selected new graphic variables; the types of visualization (maps compared, maps combined and animation) should also be taken into account. Last but not least, we should follow the conventions of individual applications and certain users.

## Acknowledgements

This work was supported by a Masaryk University as a part of research project "Analysis, evaluation, and visualization of global environmental changes in the landscape sphere" (MUNI/A/0952/2013).

## References

Hall JW, 2003, Handling uncertainty in the hydroinformatic process. Journal of Hydroinformatics, 5(4):215–231.

Harrower M, 2003, Representing uncertainty: Does it help people make better decisions. In: UCGIS Workshop: Geospatial Visualization and Knowledge Discovery Workshop.

Koláčný A, 1969, Cartographic information – a fundamental concept and term in modern cartography. The Cartographic Journal, 6(1):47–49.

Kubíček P and Šašinka Č, 2011, Thematic uncertainty visualization usability – comparison of basic methods. Annals of GIS 17(20).

Leitner M and Buttenfield BP, 2000, Guidelines for the display of attribute uncertainty. Cartography and Geographic Information Science 27:3–14.

MacEachren A et al., 1992, Visualizing uncertain information. Cartographic Perspectives, 13(3):10–19.

MacEachren A et al., 2005, Visualising geospatial information uncertainty-What we know and what we need to know. Cartography and Geographic Information Science, 32(3):139–160.

Morita T, 2004, Ubiquitous mapping in Tokyo [online]. Tokyo: ICA UPIMap. Available: http://www.ubimap.net/upimap2004/html/papers/UPIMap04-A-01-Morita.pdf

Pang A et al., 1997, Visualizing uncertainty in geo-spatial data. In: Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology. Washington, D. C.: National Academies Committee of the Computer Science and Telecommunications Board, 1-14.

Regulations (EC) No. 166/2006 of the European Parliament and of the Council of 18 January 2006 Concerning the establishment of a European pollutant release and transfer register and amending Council Directives 91/689/EEC and 96/61/EC

Senaratne HV et al., 2012, Usability of spatio temporal uncertainty visualisation methods. In: Bridging the Geographic Information Sciences. Lecture Notes in Geoinformation and Cartography, Part 1.

Zhang J and Goodchild M, 2002, Uncertainty in Geographical Information. London: Taylor and Francis.

# Improving knowledge about wildlife mobility in using geographic network analysis

E. Buard[1]

[1] COGIT Lab – IGN, 73 avenue de Paris, 94165 Saint Mandé Cedex
Email: elodie.buard@ign.fr

## 1. Context and objective

Huge amounts of data are available nowadays concerning mobility, partly because of the GPS technique improvement and increase availability. They concern different types of geographic objects such as pedestrians and cars in a city for urban planning (Nanni et al. 2013), animals for ecology (Steiniger et al. 2010) or boats and planes for navigation (Devogele et al. 2013, Baud et al. 2007). Our application concerns large African herbivores (elephants, zebras, buffaloes) moving in the Hwange National Park in Zimbabwe. We are interested in interactions between herbivores and their natural space: on the one hand, animals move according to resources availability and on the other hand, they can modify and even damage these resources in over consuming or over stamping them (Valeix et al. 2007).

In this contribution, the objective is to describe wildlife mobility, including rhythms and spatial choices of movements. In particular, we are wondering for what reasons animals choose preferentially a path rather than another.

In order to detect different rhythms in the animal movements, we automatically segment their trajectories into 2 spatio-temporal features from their GPS positions: effective movements and stops. Places of stops, called stations, are related to spatial features (water points or types of vegetation) as they are used by animals to accomplish certain activities (see section 2). This problem is similar to finding popular places like in (Benkert et al. 2010, Zheng et al. 2009). Then we spatially and semantically compare the theoretical network formed by the connected stations and the real network taken by animals constructed from the observed trajectories (see section 3). It enables to evaluate flows on each path of the networks. As a matter of fact, it raises some differences showing that animals do not choose shortest paths between stations, but the ones meeting other criteria.

## 2. Segmenting trajectories: from GPS positions to Time-Geographic concepts

10 groups of elephants, 10 of zebras and 7 of buffaloes (mean size respectively: 14, 5 100) are tracking during one year by one GPS position per hour. We assume these groups stay stable in this period as they are familial groups living together.

We want to detect places used to accomplish certain activities and then study their temporal connections. To identify these particular places, we adapt the Time-Geography concepts introduced to study human trajectories in cities (Hägerstand 1970). A place where the individual stops during certain duration is called a station. In our application, a station is a place where the group of animal stops, even in doing small movements. Stations are particularly interesting in our case since they are practiced by herbivores to collect resources and as a consequence are important features of interactions between space and animals.

To identify stations, we adopt a rule based approach depending on the speed between successive acquired positions as (Buchin et al. 2012). It brings about segmenting the trajectories of groups of animals into stations and concrete movements.

To create stations, the first step is to establish one or several speed thresholds (according to the species) between two successive positions beyond which we consider that a group does not move. This parameter has to be set up by species in regards to their maximum speeds per hour in our datasets. In our bushy study area, zebras and elephants reach 13 km/h and buffaloes 8 km/h. After a few analyses of the data and in consulting the ecologists, we fix the speed thresholds to 1% of the maximum speeds. A station, formed by several low speed successive segments, is represented as circumcircle containing them. Stations vary in durations and extents (see Figure 1), which enables to determine the activities amongst very basic ones: eat, drink and sleep. For instance, the more the station lasts, the more likely the group is to sleep.

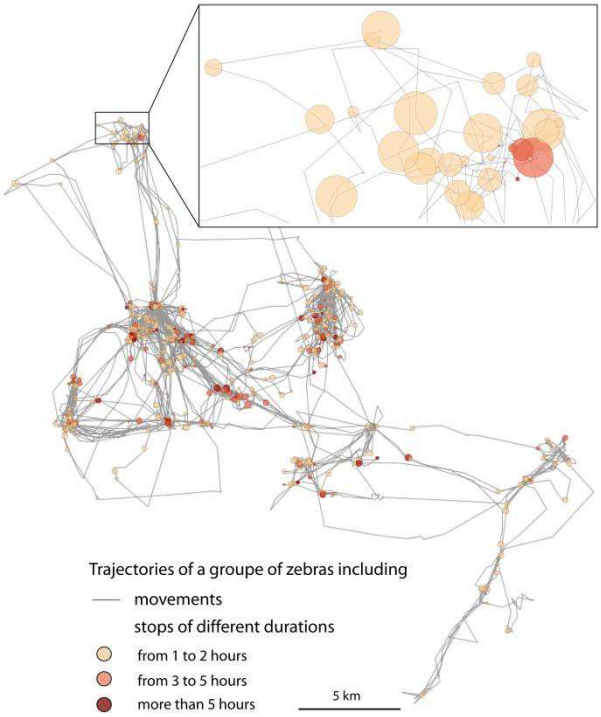



Figure 1. Durations and extents of stations

## 3. Evaluating the practiced paths: using geographic network analysis

To describe intensities of movements in stations and along paths between stations, our method relies on geographic network analysis. A geographic network is defined by its topology, nodes and edges linking the nodes. In most of the cases, the network is already set up: roads (Guo et al. 2010) or metro (Gleyze 2008). Here the animals can move freely in every direction. To construct the network, we use the previous work in considering stations as fixed nodes. Stations are first spatially clustered if they are close (less than 100m): they reflect the same spatial features, mainly water points.

Two networks are distinguished (example of construction in Figure 2) and compared:

1/ a theoretical network built by Delaunay triangulation method on stations. The Delaunay edges correspond to all the paths groups of animals could take between stations. They correspond to the shortest paths between stations, but are not necessarily practiced.

2/ a practical network according to paths effectively taken by groups.

A spatial comparison of both networks (Figure 3) shows that there are preferred paths (superimposition of paths) and not frequented paths (theoretical paths alone). It reveals that animals do not take always the shortest paths; other criteria exist. It turns out that buffaloes

prefer to move in open grassy vegetation. Moreover they can make a detour - and a longer path- to avoid dense vegetation.

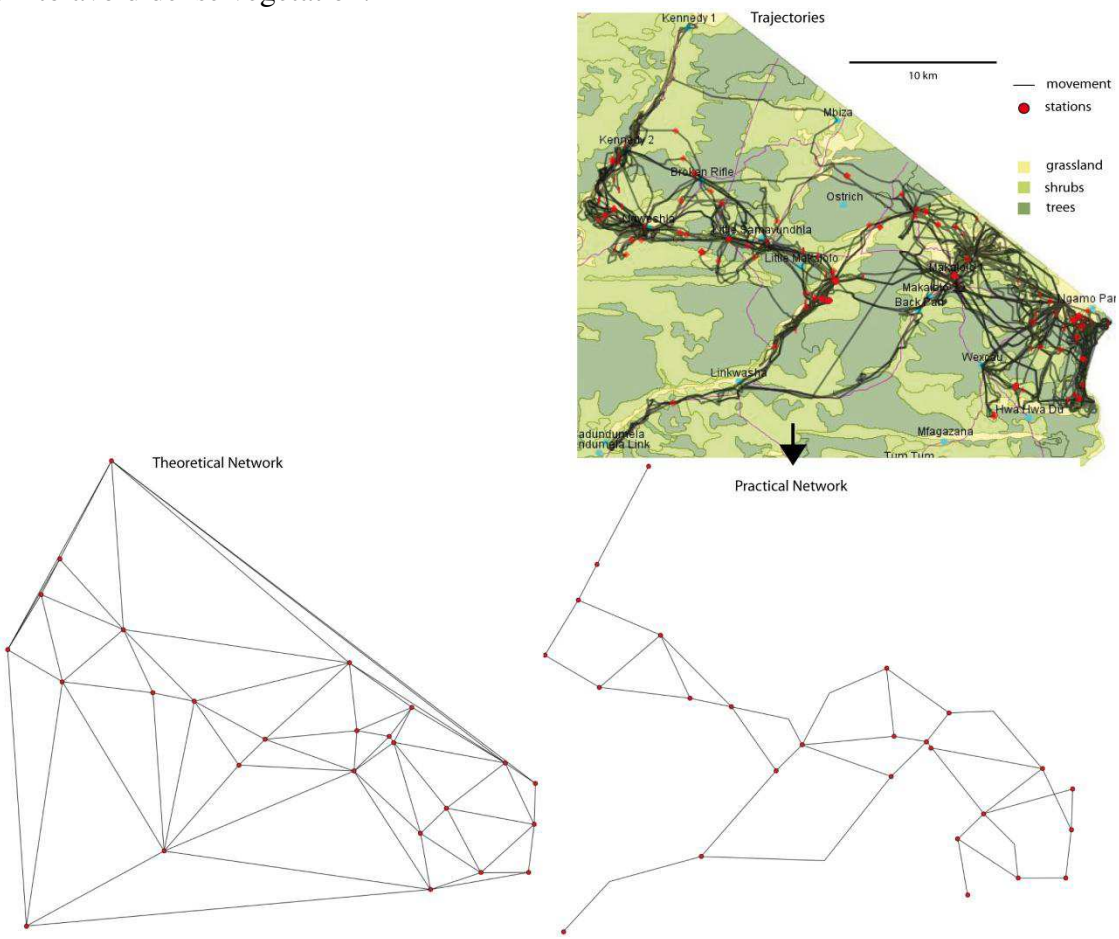



Figure 2. Two networks built from buffaloes stations

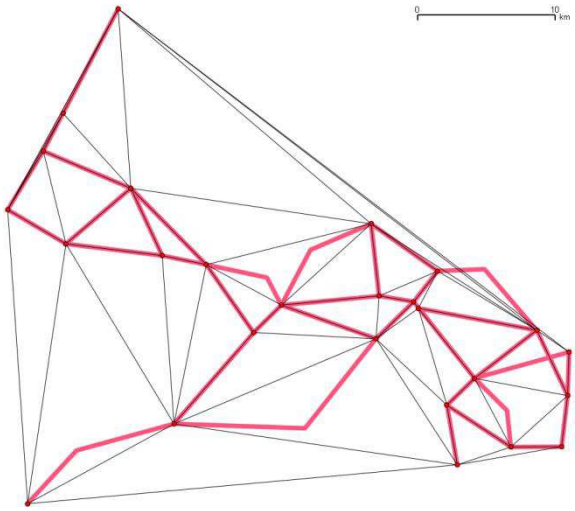


Figure 3. Spatial comparison of networks

Last, indicators are calculated and compared to evaluate different properties of stations and paths, such as: distances between stations, accessibility or centrality of stations (Mermet and Ruas 2010). Here we illustrate (Figure 4) the average accessibility of stations, which is a measure of network efficiency (Gleyze 2008). Central and peripheral nodes are equivalent

in both networks. It means practical network is enough to optimize paths. Note that this indicator matches with the observed flows: there are fewer groups in peripheral stations than in central ones.

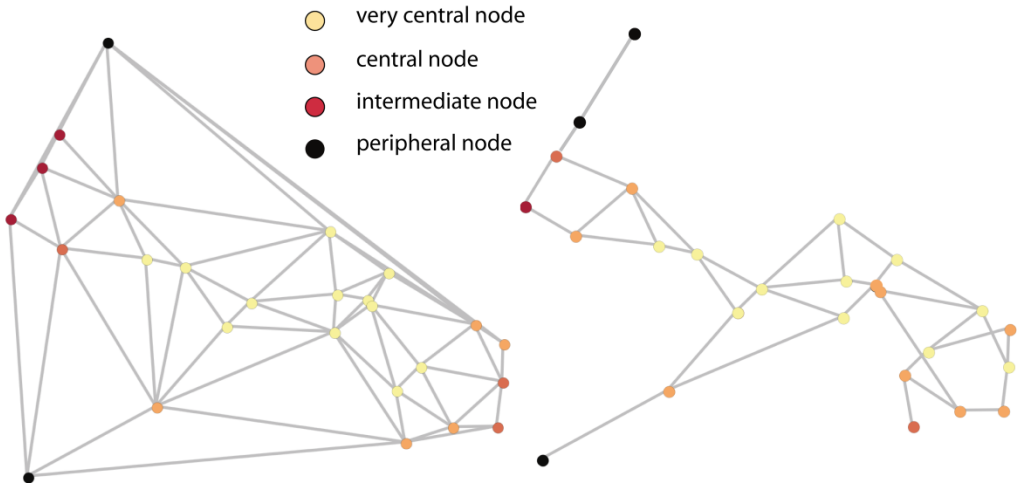



Figure 4. Average distance comparison

## 4. Conclusion

This first analysis shows that choice criteria to move between stations combine distance and type of vegetation to cross. To evaluate in which extent animals take into account vegetation, it could be interesting to estimate edge distances in function of impedance measures including vegetation.

One indicator has been highlighted here but many analyses are still to do. It will enable to evaluate the pertinence of network indicators for wildlife mobility.

## References

Baud O, El-Bied Y, Honore N and Taupin O, 2007, Trajectory comparison for civil aircraft. Aerospace Conference, IEEE, 1–9.

Benkert M., Djordjevic B., Gudmundsson J. and Wolle T., 2010, Finding Popular Places. *Int. J. Comput. Geom. Appl.* **20**, 19.

Buchin M., Kruckenberg H. and Kölzch A., 2012, Segmenting trajectories based on movement states. Proceedings of the 15th international symposium on spatial data handling,22-24 August, Bonn (Germany).

Devogele T, Etienne L and Ray C, 2013, Mobility Data: Modelling, Management, and Understanding. In: Renso C., Spaccapietra S. and Zimányi E. (eds), Maritime monitoring, Cambridge University Press, ISBN 9781107021716, Part 3, Chapter 11, 224-243.

Gleyze J.-F., 2008, Using structural approach to understand transportation networks vulnerability. European Geosciences Union, 13-18 April, Vienna (Austria).

Guo D, Liu S and Jin H, 2010, A graph-based approach to vehicle trajectory analysis. Journal of Location Based Services, 4, 183-199.

Hägerstrand T, 1970, What about people in regional science?. Papers of the Regional Science Association, 24, 1-12.

Mermet E and Ruas A, 2010, GeoGraphLab: a tool for exploring structural characteristics of transportation network. 13th International Conference on Geographic Information Science (AGILE'10), 10-14 May, Guimarães (Portugal).

Nanni M, Trasarti R, Furletti B, Renso C, Gabrielli L, Rinzivillo S and Giannotti F, 2013, Data on Science and Simulation in Transportation Research. IGI Global.

Steiniger S, Timmins TL and Hunter AJS, 2010, Implementation and comparison of home range estimators for grizzly bears in Alberta, Canada, based on GPS data. GIScience, Zurich, Switzerland

Valeix M., Chamaillé-Jammes S. and Fritz H., 2007, Interference competition and temporal niche shifts: elephants and herbivore communities at waterholes. Oecologia, 153, 3(6), 739-748.

Zheng Y., Zhang L., Xie X. and Ma W.-Y., 2009, Mining interesting locations and travel sequences from GPS trajectories. Proceedings of the 18th international conference on World wide web, April 20-24, Madrid (Spain).

# A Data Model to Capture Spatial and Temporal Exposure

Yanjia Cao[1,2,3], Chris S. Renschler[2,3], Geoffrey M. Jacquez[2]

[1]Department of Geographical and Sustainability Sciences, University of Iowa, 316 Jessup Hall, Iowa City, IA
Email: {yanjia-cao@uiowa.edu}
[2]Department of Geography, State University of New York at Buffalo, 105 Wilkeson Quad, Buffalo, NY
Email: {yanjiaca@buffalo.edu; rensch@buffalo.edu; gjacquez@buffalo.edu}
[2]LESAM Lab, State University of New York at Buffalo, 142 Wilkeson Quad, Buffalo, NY

## 1. Background and Problem Statement

Generally, according to the definition by International Programme for Chemical Safety, part of World Health Organization (WHO 2004) an exposure event is an interaction between "an agent and a target". In the field of human and environment interaction, an exposure refers to a human being impacted by different physical and psychological elements in the environment, which carries spatial and temporal information.

At ontology level, an exposure event contains exposure stressor, exposure receptor, exposure event and exposure outcome (Mattingly et al. 2012). An exposure stressor refers to a motivation in environment, especially on chemical, physical, biological and psychosocial aspects, that causes impacts on beings (Hubal 2009). An exposure receptor stands for an object or being that gets impacted by one or more environmental stressors (Mattingly et al. 2012). The interaction of a stressor and receptor is the essence of exposure science (Mattingly et al. 2012). The exposure outcome is the effect or derivative that results from the interaction between an exposure receptor and stressor during the exposure event, such as disease or a different health status (Mattingly et al. 2012). Attributes of the exposure receptors such as frequency, intensity and duration come along with characteristics of the exposure stressors such as life stages and behavior during an exposure event. Because of different stages and different distributions of stressors and receptors, an exposure event provides spatial and temporal information (Mattingly et al. 2012). With further development of exposure science, the term "exposome" was proposed by Wild (2005) to facilitate the vision of exposure measurement.

The spatial and temporal analysis is carried out in the field of Geographical Information Science (GIScience). With the integration exposure data and spatial discovery of geographical features, GIScience functionalities such as data visualization play a better role in displaying exposure situation at a geographical level (Oyana 2003). It focuses on person, time and space (Oyana 2003) rather than previously only on person and time in the field of human and environment interaction. Furthermore, the analysis within GIScience can thus provide more convincing policy constitution for the recovery from an exposure event. Generally, the spatial and temporal technique applied in exposure study is essential to understand specific problems integrated with geographical data and statistical approaches, and as well to provide reasonable judgment in problem defining and hypothesis testing (Oyana 2003). Among the statistical techniques and approaches, detecting clusters in space and time is more helpful in an exposure event analysis because it is useful in determining spatial and temporal variation and patterns of the events across a study area (Parrish et al. 2005, Parks 2008).

## 2. Results—Framework and Data Model

Based on the exposure science theory, this research designed a conceptual framework (Figure 2) for prenatal exposure under the general exposure framework. The exposure stressors are supported by the seven dimensions in PEOPLES Resilience Framework (Renschler et al., 2010). The exposure receptors specifically refer to mother and fetus, with the special life stages of three trimesters. There is also a feedback loop in this conceptual framework, which could promote recovery of prenatal exposure in a community.

Based on the research by Warren et al. (2012), the analysis of spatial and temporal data allows more easily to track specific exposure records of location and time data as well as to keep medical history of these datasets. This research developed a spatial-temporal analysis model to support an exposure event (Figure 1). The data model is built up as a conceptual work flow for the current research as well as future modification regarding the extension of the project, to retrieve quantitative and qualitative information of exposure (Nuckols et al., 2004).

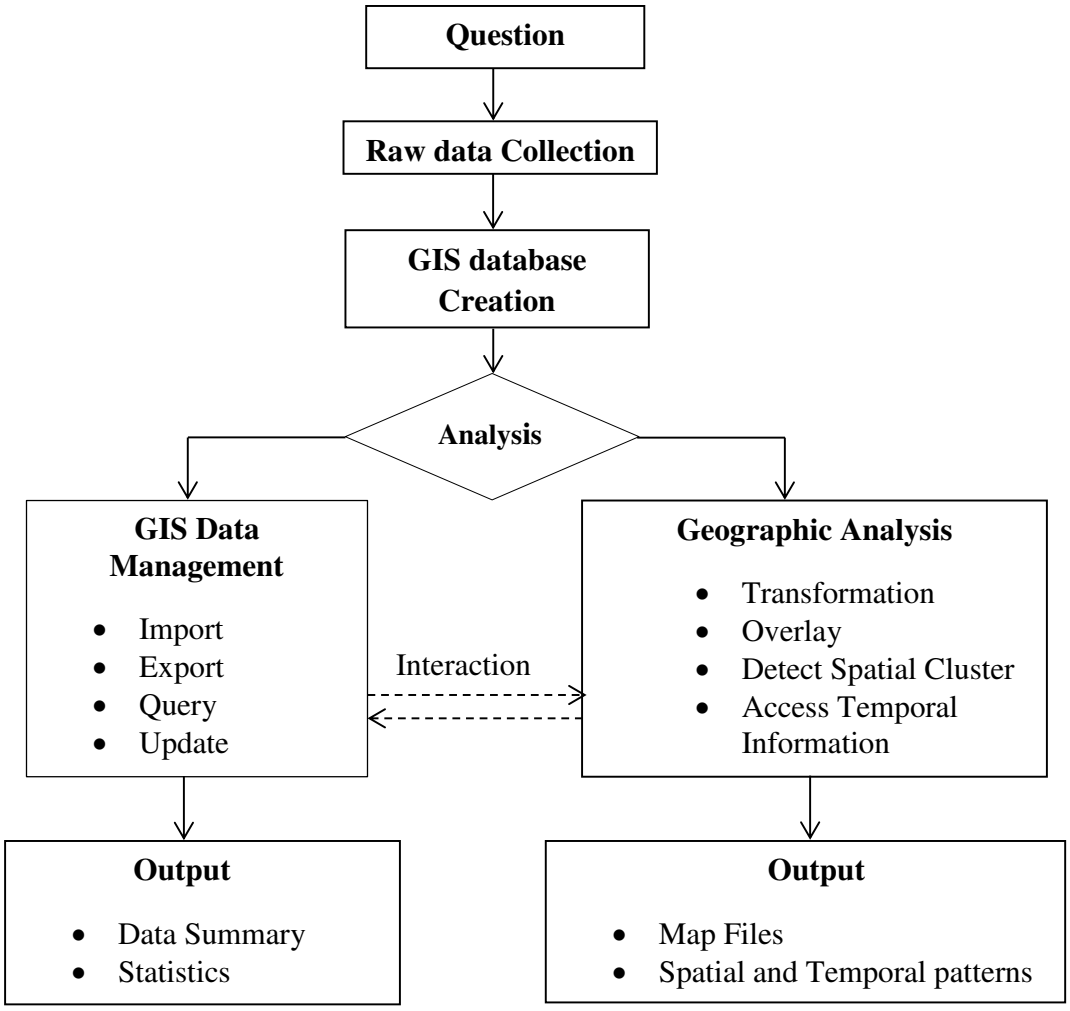



Figure 1 Data Model Design for Spatial and Temporal Analysis for Exposure

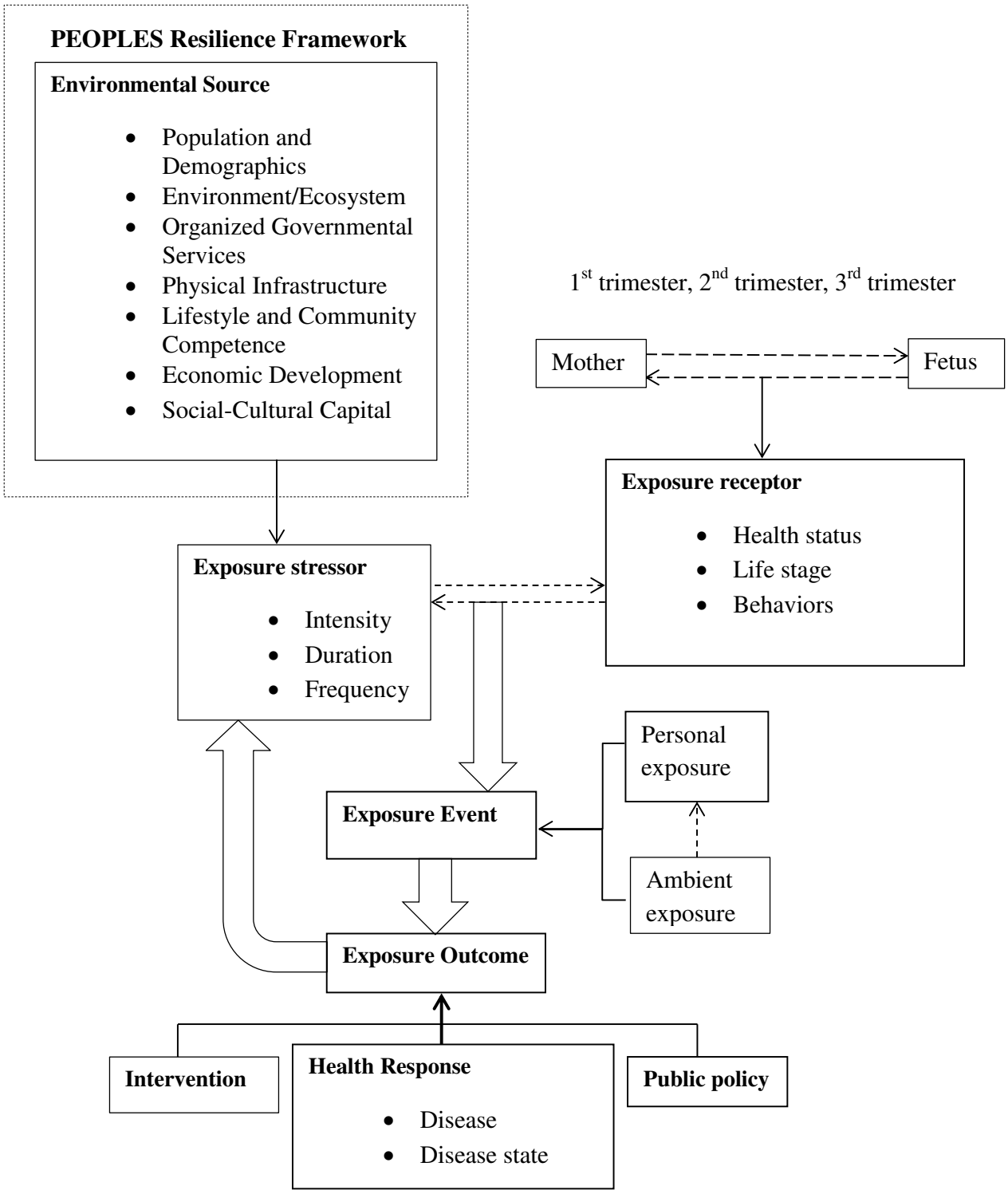



Figure 2 Conceptual Framework for Prenatal Exposure under the Larger Framework of Exposure Event

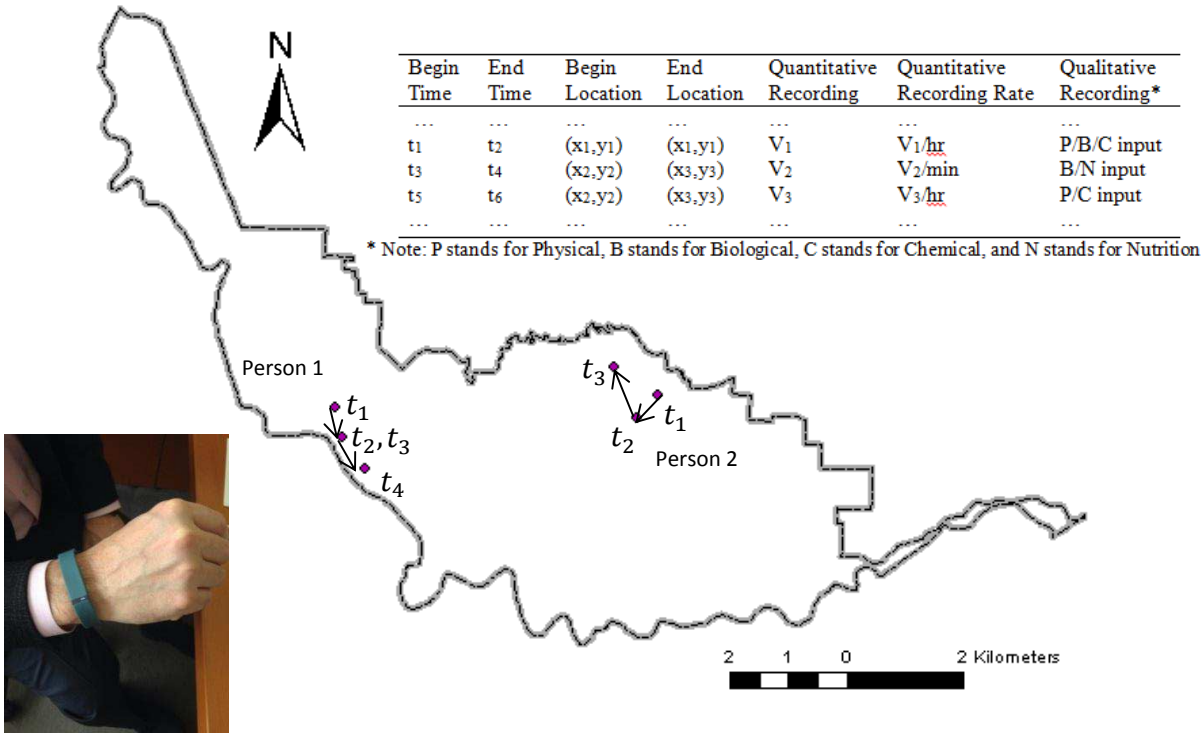
## 3. Working Example and Future Directions

The current dataset for pregnancy records are at ZIP Code Level from 2003 to 2011 in Erie County provided by New York State Department of Health Bureau of Vital Statistics (Volkman and Schoen, 2014). At the same spatial resolution level, this research chose boil water advisory records to illustrate the drinking water contamination. The spatial and temporal pattern is tracked by the following geospatial statistical methodologies (Rogerson and Yamada 2008): Local Moran's I, Join-Count Statistic, Cumulative Sum (CUSUM) and Time Series.

However, to better capture the spatial and temporal patterns of prenatal exposure in this study, this research plans to collect higher resolution dataset at individual level (Doore 2010). Shown in Figure 3, the individual data would be collected by person through a similar wearable sensor, which assists to access and record quantitative health data for a pregnant person periodically at hourly or daily increments and transmit the data via Bluetooth to mobile phones with GPS to show the location.

By collecting the accurate spatial and temporal data, the distribution of each pregnant woman can be followed and the behavior in health of each pregnant woman can be recorded. Based on this information, an enhanced research can be performed by telling the exact location of prenatal exposure. Therefore, the spatial pattern of the pregnant residents would be analyzed by Kernel Density, K-Function and nearest neighbor statistic (Rogerson and Yamada 2008).

A brief example for the new data type to be collected and analyzed is illustrated in Figure 3, the town of Gowanda. The data would be collected for each person in the format$(x_{it_j}, y_{it_j})$ to determine the location at each time period. In a similar way, information can be captured in terms of the boil water advisories, showing the exact location and time of each event and the accurate spatial range that each boil water advisory serves. Based on this information, different scenarios of exposure for each pregnant woman to drinking water can be simulated because of the variation in location in each time period within a day. Quantitative and qualitative hydration data for one person will be recorded in each time period considering the change in location. A historical record of exposure in each trimester for each pregnant woman can be therefore set up.




| Begin Time | End Time | Begin Location | End Location | Quantitative Recording | Quantitative Recording Rate | Qualitative Recording* |
|---|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... | ... |
| $t_1$ | $t_2$ | $(x_1,y_1)$ | $(x_1,y_1)$ | $V_1$ | $V_1$/hr | P/B/C input |
| $t_3$ | $t_4$ | $(x_2,y_2)$ | $(x_3,y_3)$ | $V_2$ | $V_2$/min | B/N input |
| $t_5$ | $t_6$ | $(x_2,y_2)$ | $(x_3,y_3)$ | $V_3$ | $V_3$/hr | P/C input |
| ... | ... | ... | ... | ... | ... | ... |

* Note: P stands for Physical, B stands for Biological, C stands for Chemical, and N stands for Nutrition

Figure 3 Example of Individual Data to be collected and sensor to collect data, a sample simulation in Town of Gowanda

## References


Doore, S. et al 2010. An ontology based personal exposure history. In Proceedings of the 1st ACM International Health Informatics Symposium (IHI '10), Tiffany Veinot (Ed.). ACM, New York, NY, USA, 674-683.

Hubal, Elaine A. Cohen. "Biologically relevant exposure science for 21st century toxicity testing." Toxicological sciences 111.2 (2009): 226-232.

International Programme on Chemical Safety, and Inter-Organization Programme for the Sound Management of Chemicals. IPCS risk assessment terminology. Vol. 1. World Health Organization, 2004.

Mattingly, Carolyn J., et al. "Providing the missing link: the exposure science ontology ExO." Environmental science & technology 46.6 (2012): 3046-3053.

Nuckols, John R., Mary H. Ward, and Lars Jarup. "Using geographic information systems for exposure assessment in environmental epidemiology studies." Environmental health perspectives 112.9 (2004): 1007.

Oyana, Tonny J. On the detection of patterns, trends, and distributions of respiratory diseases in western New York: an analysis of asthma using geographical information systems (GIS). Diss. State University of New York at Buffalo, 2003.

Parks, Shannon Lynn Isovitsch. Water Quality Control Through Spatial and Temporal Analysis of Water Quality Monitoring Systems. ProQuest, 2008.

Parrish, Jamie, Joanne Parkinson, and Ben Ramseth. "Advanced analysis with ArcGIS." ESRI educational services (2005).

Renschler, CS., et al. A framework for defining and measuring resilience at the community scale: The PEOPLES resilience framework. MCEER, 2010.

Rogerson, Peter, and Ikuho Yamada. Statistical detection and surveillance of geographic clusters. CRC Press, 2008.

Volkman, Diana and Schoen, Larry. Personal Communication. New York State Department of Health, Buffalo, NY (2014).

Warren, Joshua, et al. "Spatial- Temporal Modeling of the Association between Air Pollution Exposure and Preterm Birth: Identifying Critical Windows of Exposure." Biometrics 68.4 (2012): 1157-1167.

Wild, Christopher Paul. "Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology." Cancer Epidemiology Biomarkers & Prevention 14.8 (2005): 1847-1850.

# A Training-by-Example Approach for Symbol Spotting from Raster Maps

Yao-Yi Chiang, Phokgoan Chioh, Sima Moghaddam

University of Southern California, Spatial Sciences Institute, 3616 Trousdale Parkway, AHF B55
Los Angeles, CA 90089-0374
Email: {yaoyic; chioh; khashkha}@usc.edu

## 1. Introduction

Graphic symbols in maps depict important and interesting geographic phenomena, such as wetlands (Figure 1). The descriptive metadata of these symbols can be found in map labels or keys; however, labels are only capable of displaying limited information (e.g., place names) and keys provide categorical information. For example, Figure 2 shows a group of unique buildings in a U.S. Geological Survey (USGS) topographic map but the map does not provide any information about these buildings (e.g., names). Figure 3 shows a scanned map of Baghdad, Iraq where most symbols are labeled with place names but retrieving and integrating further information (e.g., addresses) of these places from other sources requires additional efforts such as using the place names and locations to search on Wikipedia or DBpedia (a structured version of Wikipedia).

In this paper, we present a training-by-example approach for spotting graphic symbols in raster maps. We demonstrate that our approach efficiently enables automatic linkages between DBpedia records and locations in a map. Traditional document analysis techniques for spotting map symbols generally require a large amount of training datasets, the presence of map keys (e.g., Samet and Soffer, 1998), or ad-hoc preprocessing steps (e.g., image thresholding) (Chiang et al, 2014; Lladós et al., 2002). In contrast, our approach takes only one user-selected symbol example to extract the locations of all symbols that have similar graphical appearance to the example.

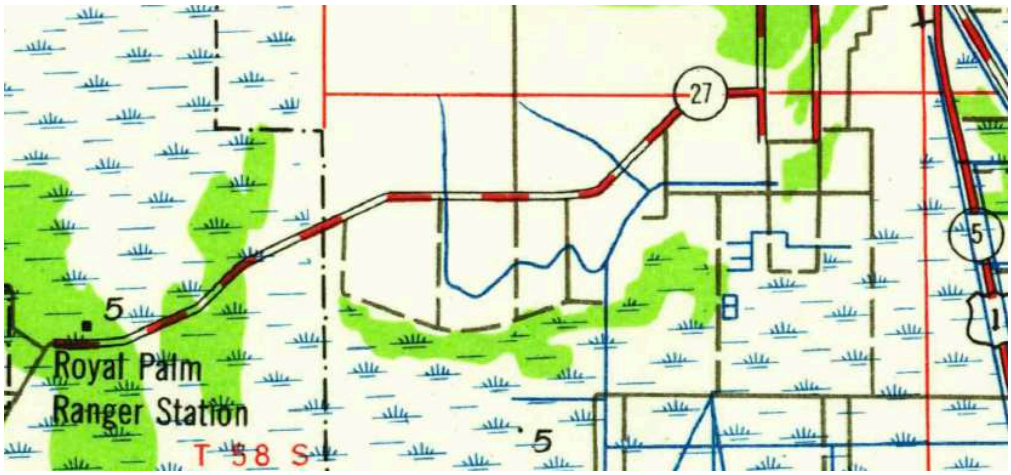



Figure 1: Wetlands in a historical USGS topographic map (Miami, Florida, circa 1958).


Figure 2: Buildings of the Park La Brea Apartment in a historical USGS topographic map (Hollywood, California, circa 1953) (left) and Google Earth imagery (right).

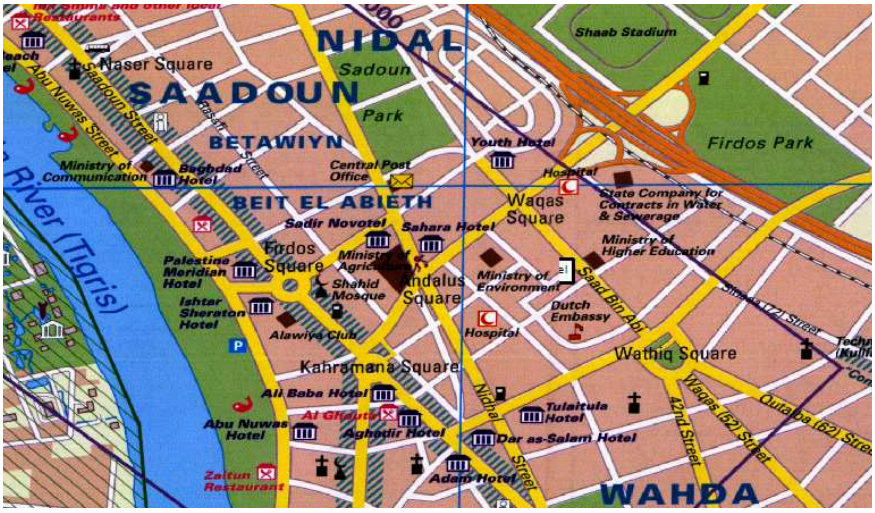



Figure 3: Symbols labeled with place names in a scanned Baghdad map.

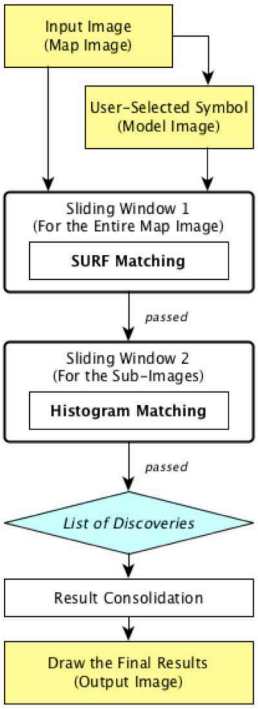



Figure 4: The SymbolRecognizer framework.

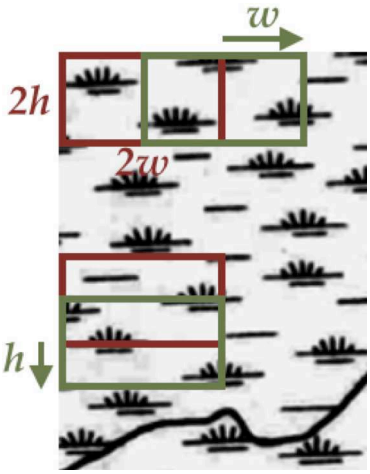



Figure 6: The SURF matching sliding window.

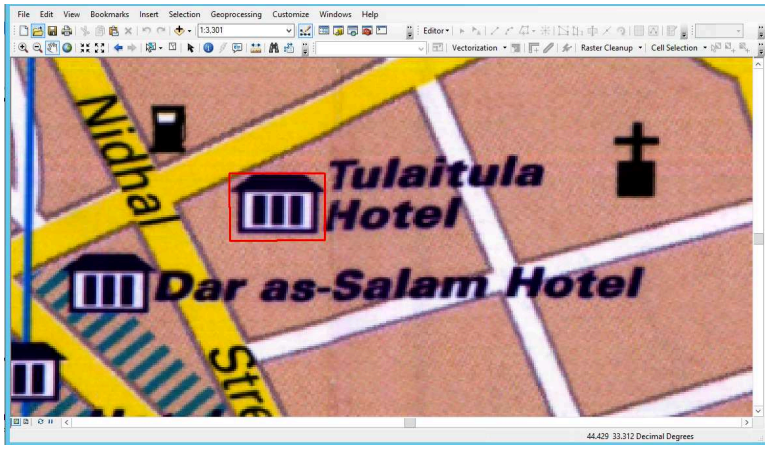



Figure 5: A user-selected symbol example.

## 2. Symbol Spotting

This section presents our symbol spotting approach called SymbolRecognizer (Figure 4). A model image is an image that covers a user-selected example in the input map (the red rectangle in Figure 5). The recognition task is to search the map for symbols that matches the model (i.e., target symbols). SymbolRecognizer utilizes a two-phase process: (1) Using the SURF (Speeded Up Robust Features) matching (Lowe, 1999; Bay et al., 2006) to efficiently identify the local regions (sub-images) where a target symbol might present and (2) Using pixel intensity distribution (with histogram matching) to verify the presence of a target symbol in each sub-image.

### 2.2 SURF (Speeded Up Robust Features) Matching

Considering a model image with width and height equal to *w* and *h* pixels, in the first phase, SymbolRecognizer uses a sliding window of the size equal to *2w* and *2h* pixels and moves *w* or *h* pixels in the horizontal or vertical direction to scan through the entire input map (Figure 6). The size of the sliding window guarantees that every target symbol is covered completely in at least one window (a sub-image). At each position of the sliding window, SymbolRecognizer detects the SURF features from the sub-image and compares the detected features with the SURF features of the model image. If the comparison result contains a high number of matched features, the sub-image is highly likely to contain a target symbol (see Lowe (1999) for details of this object recognition procedure) and is passed to the next phase.

### 2.2 Histogram Matching

The SURF matching is efficient and widely used to recognize real world objects in photography or videos, but map symbols have simpler shapes (than real world objects) and are relatively small, which can cause frequent false positives in the matching results. Therefore, SymbolRecognizer compares the pixel intensity distributions of the model image and each sub-image that passes the SURF matching to determine whether or not a target symbol presents and to extract the symbol location.

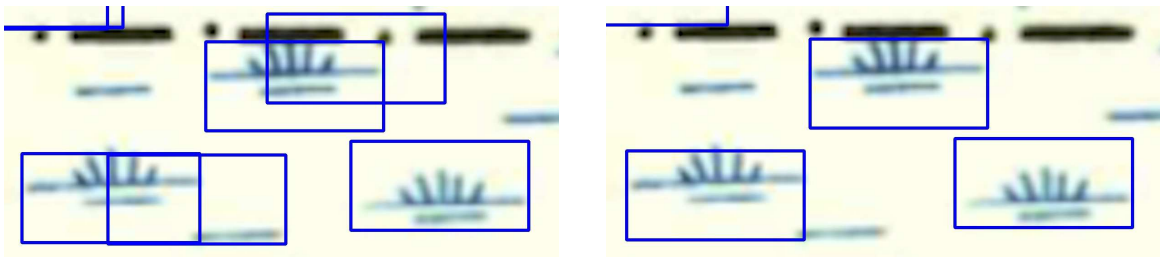
For each sub-image that passes the first phase, SymbolRecognizer uses the model image to scan from the top-left corner and moves *one* pixel in the horizontal or vertical directions (i.e., Sliding Window 2 in Figure 4). Each scanning position records a similarity score calculated using the correlation of the grayscale histogram of the model image ($H^{model}$) and the grayscale histogram of the overlapping image patch (the overlapping area between the model image and the sub-image) ($H^{patch}$). The correlation is defined as follows:

$$Similarity\ Score\ = \frac{\sum_{i=0}^{255}(H_i^{model} - \overline{H^{model}})(H_i^{patch} - \overline{H^{patch}})}{\sqrt{\sum_{i=0}^{255}(H_i^{model} - \overline{H^{model}})^2 \sum_{i=0}^{255}(H_i^{patch} - \overline{H^{patch}})^2}}$$

SymbolRecognizer uses an empirically set threshold of 90% on the similarity score to filter out the sub-images that do not contain a target symbol and to locate the symbol location. If none of the scanning positions in a sub-image has a similarity score higher than 90%, the sub-image does not contain a target symbol; otherwise, the scanning position that has the highest similarity score (in a sub-image) is the detected location of a target symbol.

### 2.3 Result Consolidation

A target symbol can be detected in overlapping sub-images during the SURF matching since the sliding window can cover a symbol more than once (Figure 7(a)). To consolidate the results, if overlapping sub-images contain multiple target symbols, SymbolRecognizer keeps only the target symbol with the highest histogram matching score (Figure 7(b)).



(a) Before consolidation. (b) After consolidation.

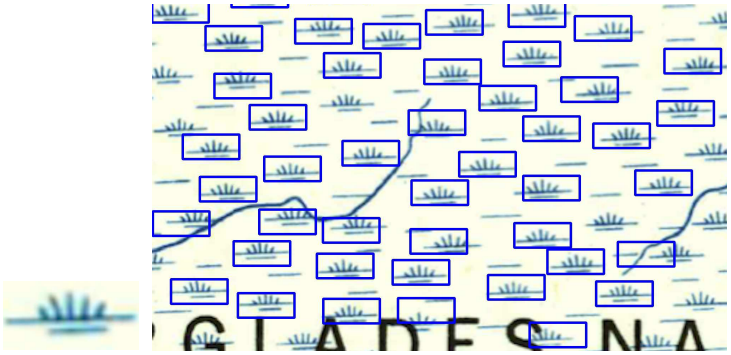
Figure 7: Result consolidation for overlapping sub-images.

## 3. Preliminary Results and Discussion

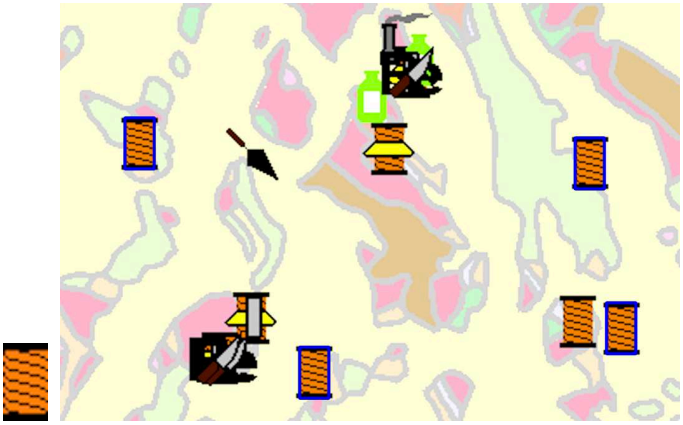
We implemented SymbolRecognizer in our map processing system, Strabo, as an Esri ArcMap plugin and tested the plugin with maps from four sources (Figure 3, Figure 8, and Table 1). For each test map, the user selected one sample symbol and Strabo automatically processed the sample to find other symbols in the map.

The results showed promising extraction precision (with only a few false positives). The USGS Hollywood map had the lowest extraction precision since the target symbols (the Park La Brea apartment buildings) are in different orientations. Although the SURF matching is rotation invariant, the histogram matching results could be compromised if the image patch did not cover the entire symbol of different orientations in the sub-image. All other test maps that contain symbols in the same orientation had more than 97% extraction precision.

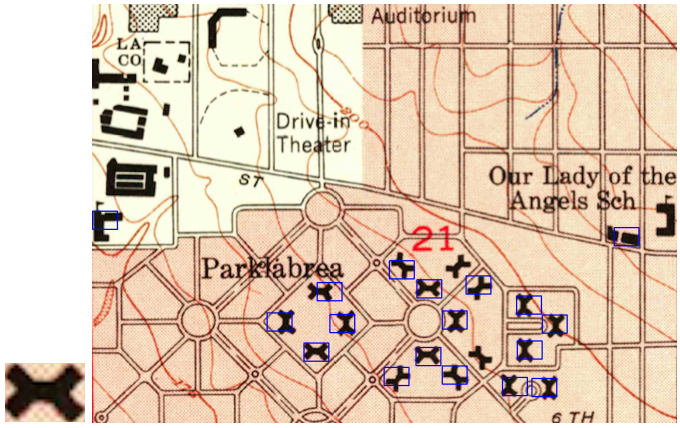
Considering the extraction recall, significantly overlapped features were the main cause of true negatives. Figure 9 shows two examples of overlapping symbols in the USGS Mine and Mineral map. The overlapping symbol to the right was detected because only a small portion of the symbol was overlapped by another symbol. The symbol to the left (Figure 9) was not detected since the entire symbol was almost covered by other symbols. The USGS Mine and Mineral map contains 12 (out of 25) significantly overlapped symbols and hence the extraction recall was the lowest among the test maps. All other test maps had more than 83% extraction recall.



(a) A historical USGS topographic map (Miami, Florida, 1958).



(b) The USGS Mine and Mineral Processing Plant Locations map.




(c) A historical USGS topographic map (Hollywood, California, circa 1953).

Figure 8: Model images (left) and sample results where blue rectangles are the recognized locations (right).

Table 1. Recognition Results.

| Source | Image Size (pixels) | # of Target Symbols | Precision | Recall |
|---|---|---|---|---|
| USGS Miami (1958) | 409x438 | 87 | 97.33% | 83.91% |
| USGS Mine and Mineral | 2465x2150 | 25 | 100% | 48% |
| USGS Hollywood (1953) | 554x396 | 18 | 88.89% | 88.89% |
| Gecko Maps, Baghdad | 5104x2616 | 17 | 100% | 88.23% |

Figure 9: Examples of overlapping symbols.

Figure 10 shows the Baghdad map with the identified symbols linked with DBpedia URIs. Once the symbols were identified, Strabo queried the DBpedia SPARQL endpoint to retrieve the nearest DBpedia entries to individual symbol locations. These entries had various DBpedia types such as Museum, Embassy, School, and Hotel. Since the identified symbols represented places in the same category, Strabo first detected the most popular category among the retrieved entries and only linked a symbol to DBpedia if the closest entry of the symbol was in the popular category. In this test area, the most popular category is Hotel and there were only four hotel entries on DBpedia.

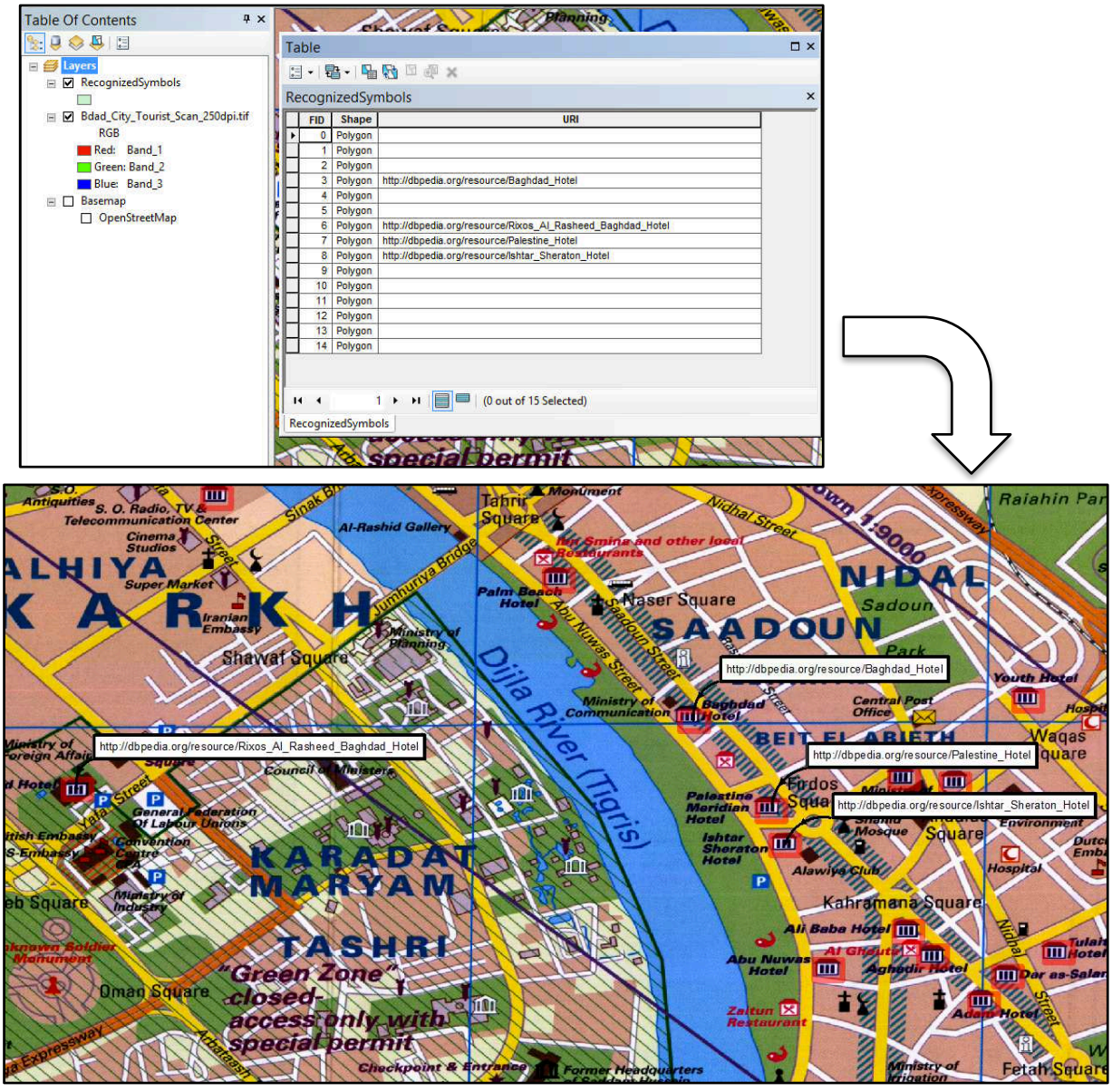



Figure 10: Automatic linkages between map locations and DBpedia records.

## 4. Summary and Outlook

We presented a training-by-example approach for symbol spotting from raster maps. Our approach requires very little user effort and can handle various types of maps and symbols. We plan to test on more symbol types and further investigate automatic methods to link the extracted symbol locations to other sources.

## References

Bay, H., Tuytelaars, T., and Gool, L. V., 2006, SURF: Speeded up robust features. In the Proceedings of the 9th *ECCV*, pages 404–417.

Chiang, Y.-Y., Leyk, S., and Knoblock, C. A., 2014, A survey of digital map processing techniques. *ACM Computing Surveys*. doi: 10.1145/2557423, in press.

Lladós, J., Valveny, E., Sánchez, G., Martí, E., 2002, Symbol recognition: Current advances and perspectives. In *GREC*, pages 104–127.

Lowe, D. G., 1999, Object recognition from local scale-invariant features. In ICCV, vol. 2, pages 1150–1157,

Samet, H. and Soffer, A., 1998, Magellan: Map acquisition of geographic labels by legend analysis. *IJDAR*, 1(2): 89–101.

# Location-allocation under conditions of limited resource: A modified Teitz and Bart approach

A.J. Comber[1], J. Dickie[1], C. Jarvis[1], M. Phillips[1], K. Tansey[1]

[1]Department of Geography, University of Leicester, Leicester, LE1 7RH, UK
Email: {ajc36; jd92; chj2; mpp2; kjt7}@le.ac.uk

## 1. Introduction

Location-allocation models seek to identify the optimal spatial arrangement of facility locations in order to satisfy some spatially distributed demand. Potential locations are evaluated using some objective and frequently this is to minimise the demand weighted distance calculated (for all demands) to the set of potential locations. Locations are commonly represented as discrete points and attributed with facility properties (supply capacity, resource volumes, etc), which may be included in the evaluation. However, in many cases, the potential selection of a particular location has immediate resource implications for regions adjacent to the set of potential locations. The implications of this situation are not considered in current algorithms. This paper therefore suggests an extension to location-allocation algorithms and applies it to the classic Teitz and Bart (1968) approach. The extension is to consider the local, spatial implications associated with any given set of potential locations. It applies this approach to determine the optimal locations for a small, community scale anaerobic digester (AD) unit.

## 2. Methods

ADs can be considered as *supply*. They require a mix of feedstocks. In this case, the problem was locate a set of hypothetical community scale ADs amongst rural locations, the *demand*. However, AD at any given supply location required agricultural and domestic feedstocks. These were to be sourced from the area around each potential AD location. Thus the critical consideration in the selection of any potential facility location was to determine how much of the surrounding area would be needed to provide feedstock resources (its catchment), and then to prevent other nearby potential supply locations in those catchments from being selected. The problem description is developed in detail below for a case study around Lincolnshire in the East Midlands region of the UK.

### 2.1 Problem Description and data

A hypothetical AD unit requiring 2.5 tonnes of dry matter per day (912.5 tonnes per year) and an optimum feedstock of 3 parts (in dry matter) of domestic food waste, 1 part cattle slurry and 1 part wheat straw, was specified. These specifications are typical of community scale AD units currently on the market and a typical feedstock. This equates to feedstocks of 547.5 t/yr of food waste, 182.5 t/yr of slurry and 182.5 t/yr wheat straw.

The supply data were generated as follows. Spatially distributed data on the annual amount of cattle slurry available were provided by the ADAS manure database (Proctor et al., 2005). This 1km resolution dataset (based on the Ordnance Survey 1km grid locations) integrates land use data agricultural census data (Comber et al., 2008) and information about local manure management practices. The agricultural census from 2010 described the amount

of land in hectares cultivated as wheat in each 1km cell. This was converted to tonnes of wheat straw per 1km cell using a factor of 3.5[1]. The third feedstock component was derived from a pycnophylactic interpolation (Tobler, 1979) of household number from the 2011 population census in each census small area (Output Area) over the same 1km grid as the agricultural data. The household number was multiplied by a factor of 0.260 to estimate the amount of household waste available as AD feedstocks in tonnes per year. The supply data are shown in Figure 1. The 'demand' was also derived from the 2011 population census using the residential population living in village areas. A demand surface was constructed by pycnophylactic interpolation of the residential population over the 1km grid.

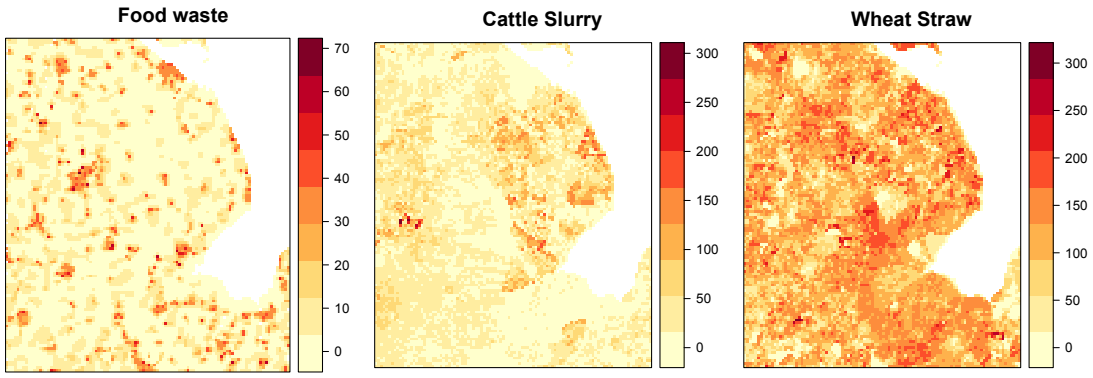



Figure 1. The spatial distribution of all AD feedstocks supply in the study area.

## 2.3 Modified p-median model

The *p*-median model (Hakimi, 1964; ReVelle and Swain, 1970) seeks to identify sets of supply locations that minimise demand weighted distances and is frequently used in accessibility analyses. It is formulated as follows:

$$\sum_i^m \sum_j^n a_i d_{ij} x_{ij} \qquad \text{(Equation 1)}$$

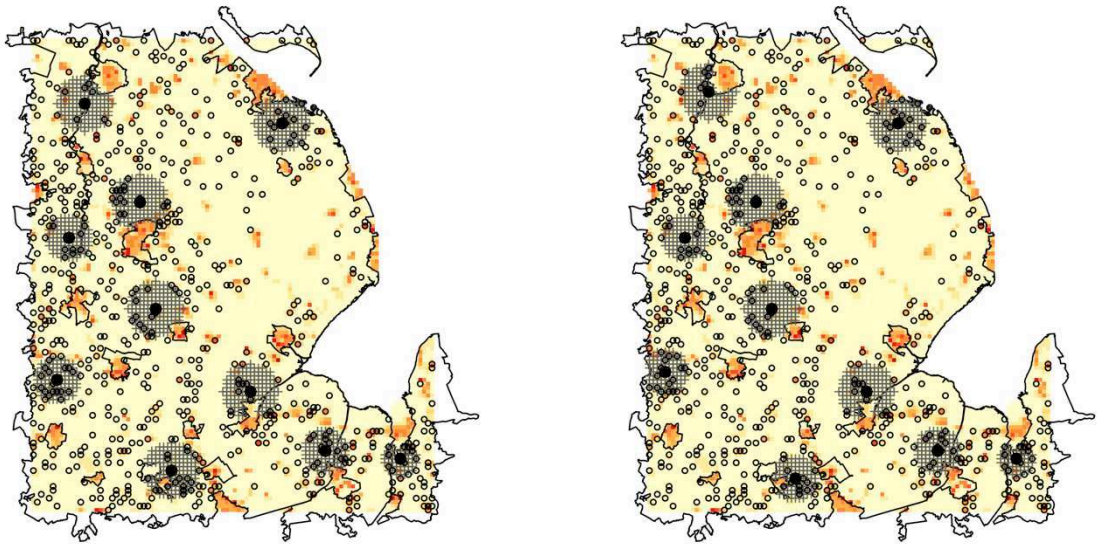
where $i$ is the index of demand locations ($1$ to $m$) and $j$ is the index of supply ($1$ to $n$), $a_i$ represents the demand at demand location $i$, $d_{ij}$ is the distance between $i$ and $j$ and $x_{ij}$ is an allocation decision variable with a value of 1 if demand at location $i$, is served by a supply $j$ and 0 if otherwise. It seeks to minimise an evaluation function. Implementations of the algorithm such as the Teitz and Bart (1968) heuristic starts with an initial set of $n$ potential supply locations and then proceeds to swap these with other locations. It then tests the new set for improvement in the evaluation function. The demand weighted distance is calculated from the sum of population weighted distances (distance multiplied by demand), where for each demand location, distance is to the nearest suggested supply location and demand is the demand weight, in this case population, at that demand location.

The problem with the classic *p*-median approach is that sufficient feedstocks for an AD are not located at each supply location. Rather they will have to be collected from farms and houses nearby. To accommodate demand locations within the catchment from the swap operation, the *p*-median algorithm was modified as follows. First, for each potential demand location being considered, all of the supply locations within the catchments were identified for each of the three feedstocks. Second, the exchange operation function was modified such that only locations outside of the current catchments were considered as potential locations.

[1] http://www.biomassenergycentre.org.uk/portal/page?_pageid=75,17972&_dad=portal&_schema=PORTAL

## 3. Results

Figure 2 shows the results of applying the new and original *p*-median models in order to select 10 and 82 locations for AD siting. There are subtle differences between the results of the two models. For the 10 location solutions, the differences are fine-drawn as both models have selected locations in similar areas. For the 82 location solutions there is greater difference in the supply areas that remain uncovered by location catchments but the difference between the models are also subtle, with what appears to be many overlapping areas in both model outputs.



a) b)

Figure 2. The 10 locations selected by a) the original p-median models and b) the p-median model modified to accommodate local catchments.

It is difficult to determine whether the locations in Figure 2 represent an improvement in resource use efficiency under the modification. However, the differences between the modified and original *p*-median models are more clearly illustrated when the unused supply locations and unused resourced are considered as the number of locations increases (Figure 3)

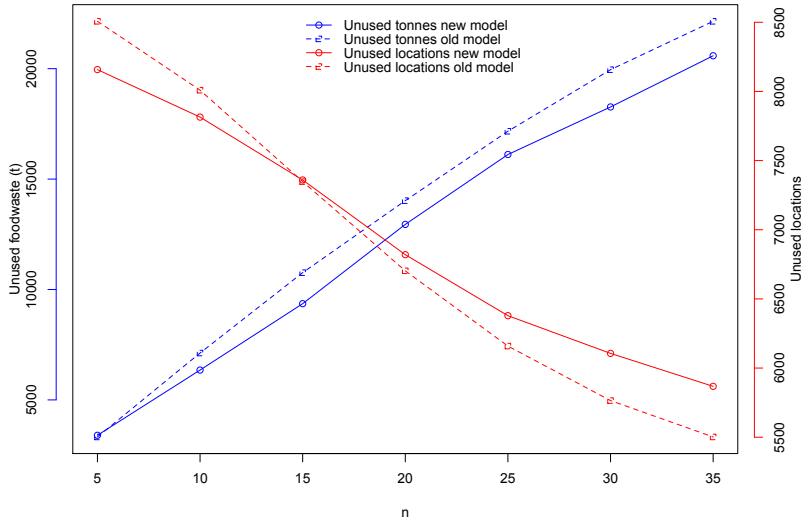


Figure 3. The unused food waste resources (tonnes and number of supply sites) selected by the classic and modified p-median model for different sized subsets of locations.

Domestic food waste was the limiting feedstock for ADs in this predominantly agricultural area. Two clear trends are evident: First that the modified p-median model consistently results in greater use of available resources than the class *p*-median model. Second that as the number of locations to be selected increases there are a greater number of unused locations. These suggest that modified model more efficiently allocates supply to demand. Future work will seek to improve the modification and will evaluate its incorporation into other heuristics such the genetic and grouping algorithms.

## Acknowledgements

The authors would like acknowledge the support of the EPSRC Grant Rural Hybrid Energy Enterprise Systems, Ref EP/J000361/1.

## References

Comber, AJ, Proctor, C and Anthony, S (2008), The creation of a national agricultural land use dataset: combining pycnophylactic interpolation with dasymetric mapping techniques. *Transactions in GIS,* 12(6): 775–791.

Hakimi, S, 1964, Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph. *Operations Research*, 12:450-459.

Procter, C., Comber, A., Anthony, S., Lyons, H. and Smith, K., (2005). Spatial data integration: the development of a manure management database for England and Wales. Pp. 565-570 in *Proceedings of the GIS Research UK 13th Annual Conference*, 6-8 April 2005 (eds Roland Billen, Jane Drummond, David Forrest, Elsa Joao), University of Glasgow

ReVelle, CS and Swain, RW, 1970, Central facilities location. *Geographic Analysis*, 2: 30-42.

Teitz, MB, and Bart P, 1968, Heuristic methods for estimating generalized vertex median of a weighted graph. *Operations Research*, 16:955-961.

Tobler WR, 1979, Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association,* 74:519–36.

# Conceptualization and Representation of Uncertainty for Science-Based Policymaking

Stephanie Deitrick[1], Joanna Merson[1]

[1]School of Geographical Sciences and Urban Planning
Arizona State University
Tempe, AZ 85287-5302
{stephanie.deitrick; Joanna.merson}@asu.edu

## 1. Introduction

Researchers and policymakers recognize the benefits of science-based policymaking for uncertain problems (NRC, 2010; Pielke, Sarewitz, and Dilling 2010). This is particularly true for complex environmental problems, such as climate change, which have escalating negative impacts on society and ecological systems. Climate change uncertainty poses a unique challenge for science-based decision support. As our understanding of climate change grows, there is still significant uncertainty about how much and how quickly the climate will change, and what impact these changes will have on society and the environment. These uncertainties vary across both geographic and temporal scales, as well as human and environmental systems (Lemos and Ramparasad 2012). The potential for climate change to have substantial socioeconomic and environmental ramifications serves as strong motivation for policymakers to obtain, understand, and incorporate climate change uncertainty into their decisions. However, even with the acknowledged need for science-based decision support, the extensive amount of climate change research available, and the willingness of policymakers to use this information, there is a disconnect between the production and use of climate change science. Policy-focused approaches to representing uncertainty can connect the understanding of uncertainty with climate-based policy decisions.

This research develops an approach for relating uncertainty and climate based policy decisions. The goal is to provide a visualization of policy outcomes that not only depicts the relationship between climate uncertainty and policy outcomes, but also provides a means to identify policies that perform at desirable levels for multiple future conditions. First, we define the relationship between science-based decision support and spatial, temporal, and attribute uncertainty. Second, we represent the relationship between policy alternatives and deep uncertainty. Lastly, we present a new method for visualizing and assessing the performance of uncertain policy outcomes.

## 2. Uncertainty Cube

In GIS, the world is modeled through a spatial filter, thus location is the usually the primary focus of representation. Some models, such as those driving geovisualizations or infographics, allow attribute values, or even temporal values, to be the primary focus of representation (Peuquet 2001). Regardless of which is the primary, they share a commonality: uncertainly is relegated to a place of least importance. It may be stored in metadata as the spatial accuracy of the original data or as error of an attribute. However, all representations for climate modeling have inherent spatial, temporal, and attribute-based uncertainty, and each element of the representation-modeling process contributes additional uncertainty. Therefore, in science-based policy decision support, the level of uncertainty should be shifted to the forefront, informing the type of decision support implemented.

As shown in Figure 1, we posit uncertainty can be conceptualized as existing simultaneously along attribute, spatial, and, temporal dimensions. Each dimension is shown along its own axis, having separate but related magnitudes of uncertainty. In this conceptualization, uncertainty refers to the degree to which the estimated value varies from the theoretical true value. Increased uncertainty is driven by standard issues of accuracy, precision, and more importantly by the conceptualizations required to represent reality in GIS form. The combination of spatial, attribute and temporal uncertainty determines the decision support approach. When there is low-uncertainly along all three dimensions, a normative approach to decision making can apply. Normative approaches focus on how people should make decisions in order to arrive at better decisions (Jonassen 2012). When there is high-uncertainty in all three dimensions, a robust approach to decision making is more appropriate. Robust decision approaches focus on identifying policies that perform well over multiple future conditions by considering multiple forms and sources of uncertainty, resulting in decisions that are less sensitive to unknowns (Walker, Rahman, and Cave 2001).

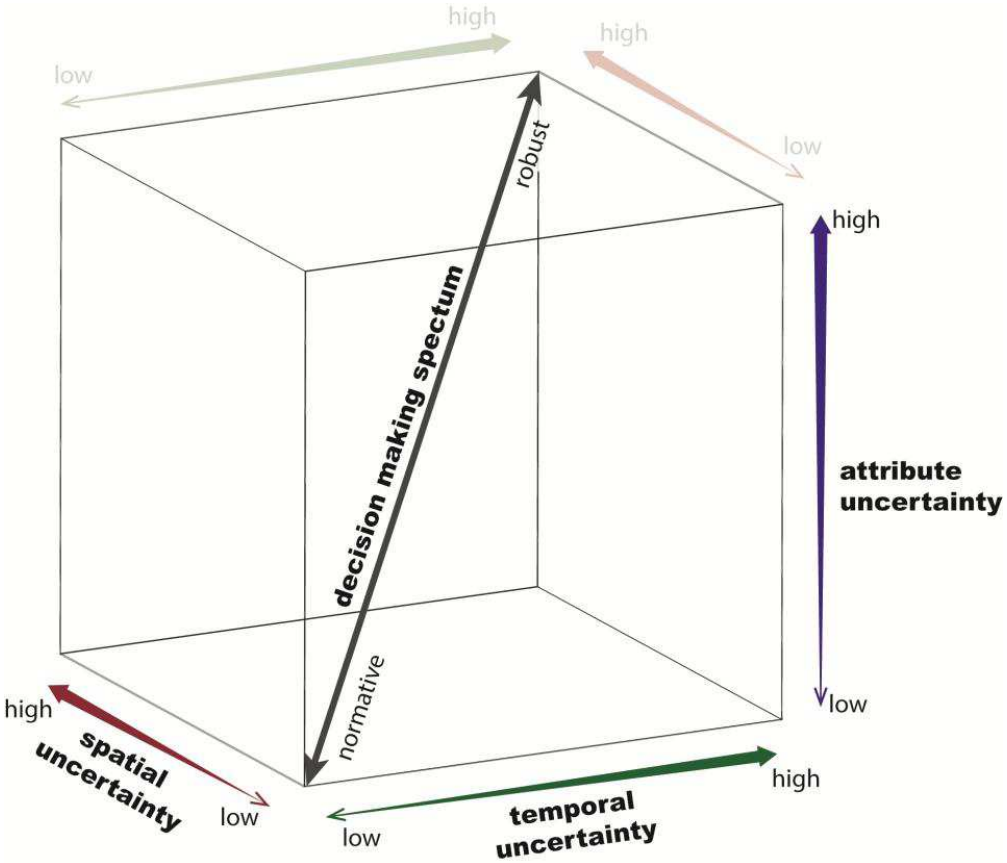



Figure 1. Conceptualization of uncertainty for science-based policymaking.

## 3. Uncertainty Continuum in GIS

An entire continuum of differing levels of uncertainty exists, ranging from total certainty and understanding at one extreme to complete ignorance at the other. Spatial, temporal, and attribute uncertainty, as discussed in the prior section, represent (theoretically) quantifiable forms of uncertainty. Deep uncertainty exists when the state of future conditions and the probability of alternatives and outcomes are unknown or cannot be agreed upon (Lempert et al 2003, Gober et. al 2010). We must also recognize that some critical uncertainties, such as those related to long term changes in physical systems or the adaptability of future societies,

are fundamentally unknowable. It is this continuum of uncertainty, where policy decisions occur in the fuzzy area between well understood and not understood, that uncertainty should influence decision support approaches.

The relationship between policy alternatives and the levels of uncertainty along the continuum is shown conceptually in Figure 2, where uncertainty exists in the attribute value estimates and the assumptions about future conditions. The combination of these uncertainties with multiple policy options results in multiple uncertainty levels that would need to be visualized in a usable manner if the projections were to be used for decision support.

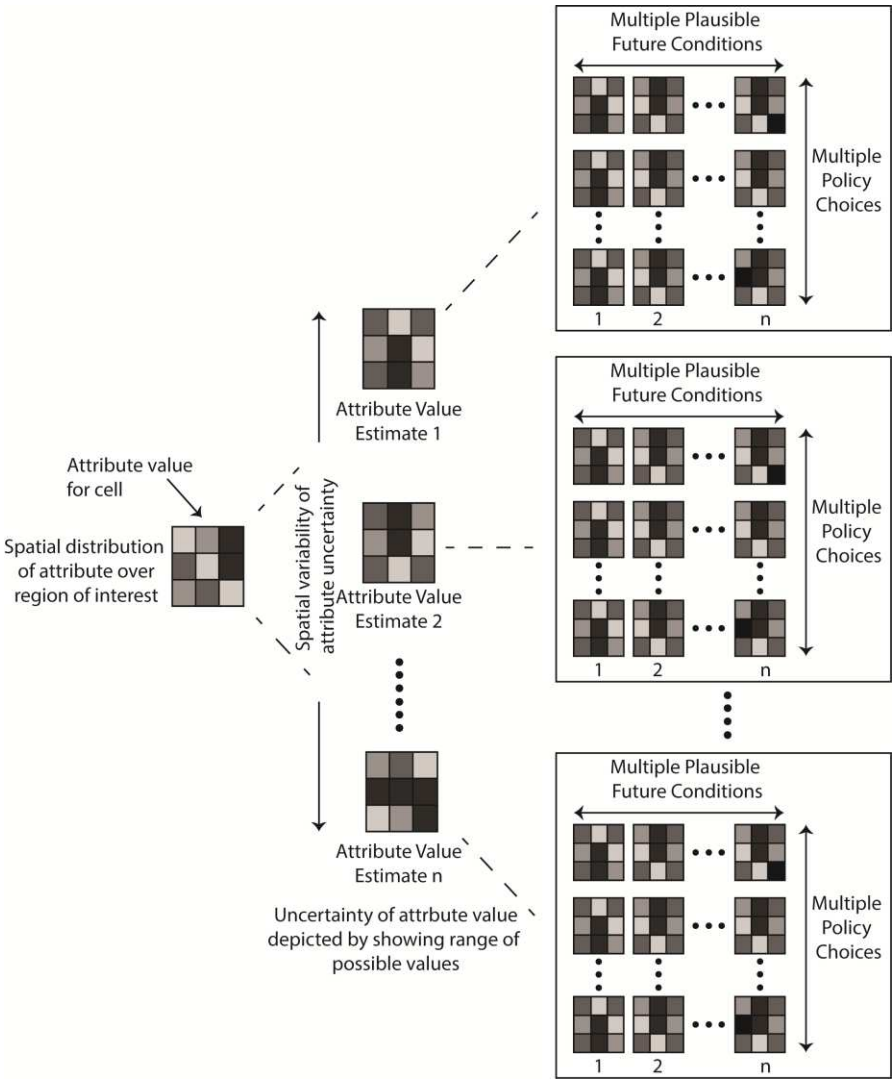


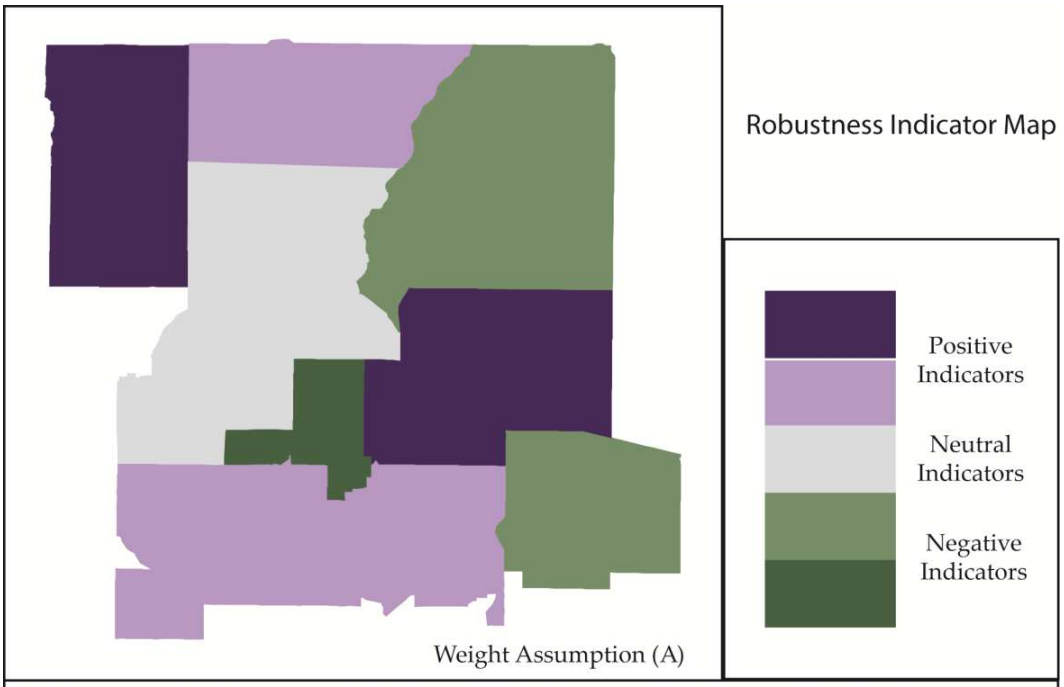

Figure 2. Conceptualization uncertainty continuum in GIS.

Since the goal of decision-making ultimately is to choose between one or more alternatives, usable methods should depict uncertainty in a manner that supports the evaluation of these alternatives. The result is that explicit quantification or visualization of specific sources, amounts or forms of uncertainty may not be possible, feasible, or desirable for deeply uncertain problems. Approaches that treat uncertainty as an inherent characteristic of decision outcomes implicitly embed uncertainty into the visualization, instead of referencing uncertainty as external to decision outcomes (Deitrick, 2013).

## 4. Robustness Indicator Maps

Deeply uncertain decision problems, or those that consist of multiple forms of uncertainty, benefit from robust decision approaches. These approaches allow decision makers to identify policies that result in outcomes that perform at a desirable level over a number of future conditions. Desirability, in this case, is measured according to criteria determined from policymaker goals or the requirements of the decision problem.

Robustness Indicator Maps (Figure 3) depict decision outcomes for a range of future conditions or uncertainties for the geographic area in question (Deitrick and Wentz, in review). The outcomes are classified based on the specifications of the decision problem and then weighted based on problem goals or guidelines. The resulting map presents a summary of how well the decision fits the criteria of the problem or the decision maker, while also reflecting the spatial variability of policy outcomes.




| Weight Assumptions (A) | Number of Outcomes in Classification | Weight of Classification | Weighted Value | Indicator |
|---|---|---|---|---|
| Sustainable 1 | 50 | 1 | 50 | 30 |
| Moderately Sustainable 0.5 | 20 | 0.5 | 10 | |
| Unsustainable -1 | 30 | -1 | -30 | |

Figure 3. Example Weights and Calculations – A sample indicator calculation is shown for each weight assumption, based on decision problem with 100 future condition outcomes (adapted from Deitrick and Wentz in review).

Figure 3 depicts indicators for a water-planning problem where water use is projected for over 100 future conditions. The indicators are calculated by first projecting water use for each future condition for each area being assessed. The projected water use is classified as sustainable, moderately sustainable, or unsustainable based on decision criteria. The total number of outcomes that fall into each category are tallied, and then weighted. In this example, decision makers identify sustainable and unsustainable water usage as equally desirable/undesirable. Moderate sustainability is seen as more desirable than unsustainable, but not as desirable as sustainable. The total number of outcomes in each classification is multiplied by the appropriate weights to obtain the indictor value for the geographic area.

Positive indicators show areas that result in overall sustainable water use, while negative indicators reflect overall unsustainable water use. The indicator map summarizes how well decision outcomes fit the decision goals.

## 5. Conclusions

The work presented here illustrates how multiple levels of uncertainty can be integrated into GIS analysis and results for policy decision support. In policy decision problems where uncertainty is a key consideration, shifting from a spatial to uncertainty focus, opens up approaches for representing deeply uncertain problems in a useful and usable manner.

## 6. References

Deitrick, S, 2013, Uncertain Decisions and Continuous Spaces: Outcomes Spaces and Uncertainty Visualization. In *Understanding Different Geographies, Lecture Notes in Geoinformation and Cartography*, ed. K. Kriz, W. Cartwright, and M. Kinberger, 117-134. Berlin Heidelberg: Springer-Verlag.

Deitrick, S and Wentz E, In Review, Developing implicit uncertainty visualization methods motivated by theories in decision science. *Annals of the Association of American Geographers*.

Gober P, Kirkwood CW, Balling RC, Ellis AW, and Deitrick S, 2010, Water Planning Under Climatic Uncertainty in Phoenix: Why We Need a New Paradigm. *Annals of the Association of American Geographers*, 100(2): 356-372.

Jonassen, D, 2012, Designing for decision-making. Educational Technology Research and Development 60 (2): 341-359.

Lemos MC, Kirchhoff CJ, and Ramparasad V, 2012, Narrowing the climate information usability gap. *Nature Climate Change*, 2:789–94.

Lempert, RJ, Popper S, and Bankes S, 2003, *Shaping the Next One Hundred Years: New Methods for Quantitative, Long Term Policy Analysis*. Report MR-1626-RPC; RAND, Santa Monica, CA, USA.

National Research Council, 2010, America's Climate Choices: Panel on Advancing the Science of Climate Change. The National Academies Press, Washington, DC.

Peuquet DJ, 2001, Making space for time: Issues in space-time data representation. *GeoInformatica*, 5(1): 11-32.

Pielke R, Sarewitz D, and Dilling L, 2010, *Usable Science: A Handbook for Science Policy Decision Makers*. Washington, DC.

Walker WE, Rahman SA and Cave J, 2001, Adaptive policies, policy analysis, and policy-making. *European Journal of Operational Research* 128 (2): 282-289.

# Comparing Terrain Categories in Wikipedia for Spatial Data Integration in CyberGIS

Chen-Chieh Feng[1] and Alexandre Sorokine[2]

[1]Department of Geography, National University of Singapore, AS2, 1 Arts Link, Singapore 117570
Email: geofcc@nus.edu.sg

[2] Computational Science and Engineering Division, Oak Ridge National Laboratory, PO Box 2008 MS 6017, Oak Ridge, TN 37831
Email: sorokina@ornl.gov

## 1. Introduction

The paper explores the potential use of Wikipedia in spatial data infrastructure research. Its overall aims are to understand the meanings of terrain categories (i.e., kinds of terrains) as understood by the contributors of Wikipedia and the multilingual compatibility of related terrain categories in Wikipedia. The study is motivated by the recent linked data movement in geospatial domain where information on Wikipedia, or its derivative product (e.g., DBpedia or YAGO2), are often used to tag user-generated spatial data sets for automatic data access and integration over the Internet (Hellmann and Auer 2013). By understanding the meanings of terrain categories in Wikipedia, the study reduces the possibility to integrate semantically incompatible data sets. The study is further motivated by the ethonophysiography hypothesis, which states that languages and cultures have bearings on the conceptualizations of landscape categories (Mark et al 2011). As such, different languages may or may not have similar terrain categories. Understanding how related terrain categories in two languages correspond to each other contributes to more accurate handling of multilingual data search and discovery in spatial data infrastructures.

Existing work comparing the meanings of terrain categories in digital geospatial databases have been based on authoritative data sources (Feng and Sorokine, forthcoming, Kavouras et al, 2005) or by using the generic part of each place name as a proxy for defining terrain categories (Feng and Mark 2012, Derungs et al, 2013). On the contrary, Wikipedia is one source of user-generated content where terrain categories are created by anyone with access to the Internet. Among other sources of user-generated content, such as folksonomies, GeoNames, and OpenStreetMap, Wikipedia provides richer information for defining terrain categories and it provides entries in all major languages. In this paper we examine terrain categories from English, Mandarin, and Russian Wikipedia pages and provide preliminary comparison on the types of terrain categories and structures between terrain categories from Wikipedia pages.

## 2. Data Sources and Tools

### 2.1 Structure of Wikipedia

Wikipedia has several means to structure its content. Here, two such constructs relevant to this paper are described. First, each Wikipedia page is referenced to one or more “Wikipedia Categories” – groups of articles on related topics designated by the writers or editors of the page. Wikipedia Categories allow a Wikipedia user to browse articles of the same or related topics. The articles themselves describe kinds, instances, or contain other information. Given that the term category in this paper refers to kinds, we use “Wikipedia Category” to refer to

Wikipedia grouping of pages into related topics. Second, Wikipedia is organized in a way that permits multiple related articles to be connected. For some articles, Wikipedia offers templates (standardized fragments of the pages that can be used in other articles) and standardized infoboxes to provide structured summary of the information on the pages.

### 2.2 Extracting Terrain Categories from Wikipedia Articles

The extraction of terrain categories from Wikipedia involved the following steps. First, we used the list of terrain categories in the spatial data standards of the US, Russian Federation, and Taiwan as seed categories to iterate through Wikipedia pages. Second, for consistency purpose we used a set of scripts to automatically retrieve the Wikipedia pages of these terrain categories. To ensure only quality pages were retrieved, we used pages with longer duration of existence and higher number of edits. The resulting list was then cleaned up by keeping only the relevant Wikipedia entries. Third, the previous two steps were repeated until no new relevant Wikipedia articles can be retrieved. After the lists of terrain categories have been created for the three languages we compared the structure, top-level Wikipedia Categories, defined as the most generic categories for organizing Wikipedia pages of terrain categories, and the relations between related terrain categories.

## 3. Preliminary Findings

### 3.1 Wikipedia Categories

The total numbers of terrain categories from Wikipedia, compared to that in the data standards we used, range from significantly larger in Mandarin and in Russian to about the same in English. Breaking down the total numbers by kinds of landforms, Wikipedia in each language has a unique distribution. More Wikipedia Categories are found in fluvial, glacial, and erosion landforms for English, mountain, island, and watercourse-related landforms for Mandarin, and glacial landforms and channels for Russian. Note that in the national data standards of Taiwan and Russian Federation, there are also larger numbers of watercourse- and channel-related categories, respectively. Similar pattern was observed for fluvial and erosional landforms in English, but Wikipedia has roughly twice the number categories than the data standard we consulted.

Certain category types are much more prominent in data standards or in Wikipedia. For Russian, categories for boundary features are abundant in data standards but non-existent in Wikipedia. For Mandarin, categories for island- and mountain-related landforms in data standard are outnumbered by that in Wikipedia.

Top-level Wikipedia Categories of the three languages are similar; all of them have mountain, river, fluvial, glacial, oceanic, Aeolian, and slope landforms as top-level Wikipedia Categories and separate Wikipedia Categories by forms and processes. Note that coastal landform is also a top-level category in both English and Mandarin Wikipedia, but it is a sub-category of hydrology in Russian Wikipedia. Unique to the Mandarin Wikipedia pages is the set of Wikipedia Categories that explicitly recognize regions at multiple elevations ranging from mountaintop to seabed, and shapes.

### 3.2 Relations between Wikipedia Categories

Relations between Wikipedia Categories generally hold is-a relations given that Wikipedia Categories discussed in this paper exclude entities other than genuine categories. Such observations does not imply non-existence of other relations; many other well-known formal

relations can be found, including but not limited to part of, located in, spatial coincidence and overlap, aggregation, and contributing process.

As indicated in Section 2.1, a Wikipedia page is referenced to one or more Wikipedia Category. For English terrain categories in Wikipedia, multiple references to Wikipedia Category can enrich the descriptions of terrain categories. For example, plateau, which is referenced to both mountainous and slope landforms, is semantically richer than floodplain, which is referenced to fluvial landforms only, because of the one additional relation specified for plateau. Multiple references to Wikipedia Category for Mandarin terrain categories in Wikipedia, on the contrary, are seldom found despite that the terms of several terrain categories suggest their relations to both forms and processes. Russian Wikipedia pages that describe terrain categories typically contain lots of textual information with few or none relevant infoboxes and they link to at most two Wikipedia Categories. Textual description may include lists of subcategories and lists of morphological parts of the geographic features but the relations in such taxonomies and partonomies are not automatically parsable.

## 4. Conclusion

The paper examined Wikipedia pages in English, Mandarin, and Russian for the types of terrain categories, the relations between these terrain categories and explored their commonality for understanding the implications of using Wikipedia to enable semantic spatial data integration over the Internet. Current approach of analysis relies on some of the structured information in Wikipedia described in Section 2.1. In the future, it would be worth exploring how other Wikipedia constructs, e.g., multilingual links, can be used to understand the conceptualizations of terrain categories by the contributors of Wikipedia. We will also glean through the definitions and descriptions of terrain categories and explore methods to systematically consume these data for integrating multilingual spatial data sets.

## References

Derungs C, Wartmann F, Purves RS, and Mark DM, 2013, The meanings of the generic parts of toponyms: Use and limitations of gazetteers in studies of landscape terms. In T Tenbrink, J Stell, A Galton, and Z Wood (eds), Spatial Information Theory, Springer, Berlin, 261-278.

Feng C-C and Mark DM, 2012, Exploring NGA GEOnet Names Server Data for cross-linguistic research on landscape categories: A case study of some toponyms in Malaysia and Indonesia. GIScience 2012 Extended Abstract.

Feng C-C and Sorokine A, forthcoming, Comparing English, Mandarin, and Russian hydrographic and terrain categories. International Journal of Geographical Information Science, DOI: 10.1080/13658816.2013. 831420.

Hellmann S and Auer S, 2013, Towards web-scale collaborative knowledge extraction. In I Gurevych and I Kim (eds), The People*'s Web Meets NLP* – Collaboratively Constructed Language Resources. Springer-Verlag, Berlin Heidelberg, 287-314.

Kavouras M, Kokla M, and Tomai E, 2005, Compare categories among geographic ontologies. Computer and Geosciences, 31(2), 145-154.

Mark DM, Turk A G, Burenhult N and Stea D, 2011, Landscape in language: An introduction. In D M Mark, A G Turk, N Burenhut and D Stea (eds), Landscape in Language – Transdisciplinary Perspectives. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1-24.

# Strong Spatial Cognition

Christian Freksa

SFB/TR 8 Spatial Cognition, University of Bremen
Enrique-Schmidt-Str. 5, 28359 Bremen, Germany
Email: freksa@uni-bremen.de

## 1. Introduction

The ability to solve spatial tasks is crucial for everyday life and thus of great importance for cognitive agents. A common approach to modeling this ability in artificial intelligence (AI) has been to represent spatial configurations and spatial tasks in form of *knowledge about* space and time. Augmented by appropriate algorithms such representations permit the computation of knowledge-based solutions to spatial problems. In comparison, natural embodied and situated cognitive agents often solve spatial tasks without detailed knowledge about underlying geometric and mechanical laws and relationships; they directly relate actions and their effects due to spatio-temporal affordances inherent in their bodies and their environments. We argue that spatial and temporal structures *in the body and the environment* can substantially support (or even replace) reasoning effort in computational processes. This principle is applied, for example, in descriptive geometry for geometric problem solving, but has not been investigated as a paradigm of cognitive processing. The relevance of this principle may not only be to overcome the need for detailed knowledge for a knowledge-based approach, but also to understand the efficiency of natural problem solving approaches. We use the term *Strong Spatial Cognition* in analogy to Searle's notion of *Strong AI* (Searle 1980) to signify that we pursue the construction of an embodied and situated cognitive system as compared to a simulation of its behavior in a purely knowledge-based system.

## 2. Architecture of Cognitive Systems

Cognitive agents such as humans, animals, and autonomous robots comprise brains (resp. computers) connected to sensors and actuators. These are arranged in their (species-specific) bodies to interact with their (species-typical) environments. All of these components need to be well tuned to one another to function in a fully effective manner. For this reason, it is appropriate to view the entire aggregate (cognitive agent including body and environment) as a 'full cognitive system' (Fig. 1).

The present work investigates the distribution, coordination, and execution of tasks among the system components of embodied and situated spatial cognitive agents. From a classical information processing / artificial intelligence point of view, the relevant components outside the brain or computer would be formalized in some knowledge representation language in order to allow the computer to perform formal reasoning or other computational processing. Physical, topological, and geometric relations are transformed into abstract *information* about these relations and the tasks are then performed entirely on the information processing level, where true physical, topological, and geometric relations no longer persist.

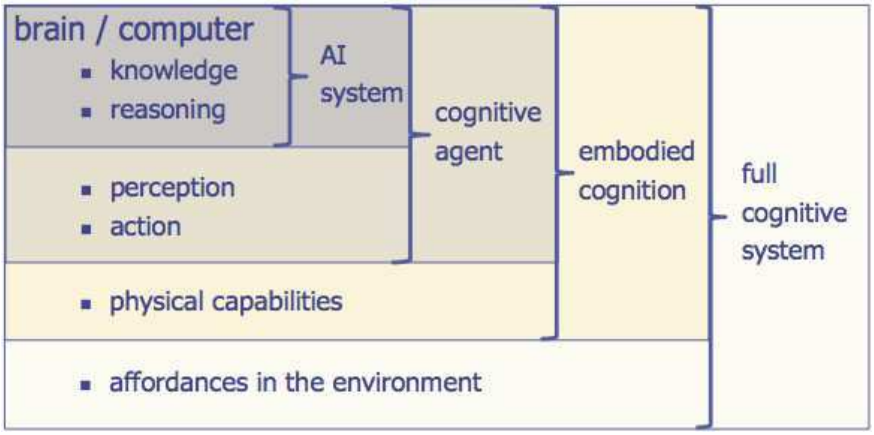



Fig. 1 Structure of a full cognitive system

However, this classical information processing-oriented division between brain/computer on one hand and perception, action, body, and environment on the other hand is only one way of distributing the activities involved in cognitive processing. Alternative ways would be (1) to maintain some of the spatial relations in their original form or (2) to use only 'mild abstraction' for their representation. The best-known example of mild abstraction is geographic maps: here certain spatial features are represented by identical spatial features (e.g. distances by distances, angles by angles, and shapes by shapes – i.e. in their original form); only few relations are abstracted by transformation (e.g. distances and sizes could be scaled). As a result, physical operations such as perception, route-following with a finger, and manipulation (e.g. perspective change) may remain enabled as in the original domain.

Maintaining relations in their original form corresponds to what Norman (1980) named *knowledge in the world.* Use of knowledge in the world requires perception of the world to solve a problem. While the need of perception may appear as a disadvantage in comparison to computational symbol processing on fully abstractly represented knowledge it has the advantage of enabling direct use of our experiential abilities of dealing with space, even if we lack the analytical knowledge about the properties of space that is required for employing abstract representations. Perception also is required for the use of mildly abstracted representations – but the perception task can be easier than the same task under real-world conditions, for example due to the modified scale that allows for substituting body movements in the geographic world by finger movements and / or eye movements in the map representation.

A main research hypothesis for studying physical operations and processes in spatial and temporal form in comparison to formal or computational structures is that spatial and temporal structures *in the body and the environment* can substantially support reasoning effort in computational processes. When comparing the use of such different forms of representation (formal, mild abstraction, original) we see that the processing structures of problem solving processes differ (Marr 1982). Different structures facilitate different ease of processing (Sloman 1985).

This hypothesis can be plainly formulated as:

***manipulation + perception simplify computation***

While the underlying concept is well known – for example, it is applied in descriptive geometry for geometric problem solving when we construct geometric entities by physical use of ruler and compass – it has not been investigated as a principle of cognitive processing.

Abstract reasoning about the world can be considered the most advanced level of cognitive ability; this ability requires a comprehensive understanding of the mechanisms responsible for the behavior of bodies and environments. But many natural cognitive agents (including adults, children, and animals) lack a detailed understanding of their environments and still are able to interact with them rather intelligently. For example, they may be able to open and close doors in a goal-directed fashion without understanding the mechanisms of the doors or locks on a functional level. This suggests that knowledge-based reasoning may not be the only way to implementing problem solving in cognitive systems.

Alternative models of perceiving and moving goal-oriented autonomous systems have been proposed in biocybernetics and AI research to model aspects of cognitive agents (e.g. Braitenberg 1984, Brooks 1991, Pfeifer and Scheier 2001). These models physically implement perceptual and cognitive mechanisms rather than describing them formally and coding them in software. Such systems are capable of intelligently dealing with their environments without encoding knowledge about the mechanisms behind the actions. The background of the present work has been discussed in detail in (Freksa 2013, Freksa and Schultheis in press).

## 3. Approach

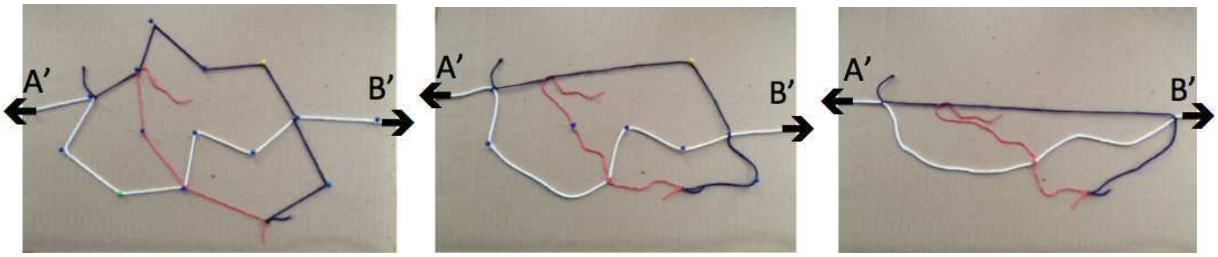
With this work, we go an important step beyond previous embodied cognition approaches to spatial problem solving. The proposed paradigm shift not only aims at preserving spatial structure, but also will make use of identity preservation; thus, spatial objects and configurations will be represented by themselves or by *physical spatial* models of themselves, rather than by their abstract representations. In this way we can avoid loss of information due to early representational commitments: we do not have to decide prematurely which aspects of the world to represent and which aspects to abstract from; this can be decided partly during the problem solving process. At that stage, additional contextual information may become available that can guide the choice of the specific representation to be used.

Perhaps more importantly, physical objects and configurations usually are spatially aggregated in a natural and meaningful way. For example, a chair may consist of a seat, several legs, and a back; if I move one component of a chair, I automatically (and simultaneously!) move the other components and the entire chair, and vice versa. This property is essential for our experience and use of spatial objects; it is not intrinsically given in abstract representations of physical objects. In contrast, if I manipulate symbolic *representations* of a chair in order to infer spatial implications of physical manipulations, I have to process knowledge about the chair's components before I can infer implications on the whole chair (or vice versa).

Thus, from an information processing perspective, there is a big difference in the characteristics of the processes in the problem domain and in its representation. From a cognitive point of view, extending information processes by physical manipulation and perception of spatial entities may be a very useful feature, as no computational processing cycles are required for simulating physical effects or for reasoning about them. Thus, manipulability of physical structures may become an important feature of cognitive processing, and not merely a property of physical objects.

Similarly, we aim at dealing with perception dynamically, for example allowing for "on-the-fly" creation of suitable spatial reference frames: by making direct use of spatial configurations, we can avoid deciding *a priori* for a specific spatial reference system in which to perceive a configuration. As we know from problem solving in geometry and from spatial cognition, certain reference frames may allow a spatial problem to collapse in dimensionality and difficulty. For

example, determining the shortest route between two points on a map boils down to a 1-dimensional problem (Dewdney 1988). However, it may be difficult or impossible to algorithmically determine a reference frame that reduces the shortest route finding task given on a 2- or 3-dimensional map to a 1-dimensional problem. A spatial reconfiguration approach that makes use of the physical affordance 'shortcut', easily reduces the shortest route problem from 3D or 2D to 1D, as depicted in Fig. 2.




(a) (b) (c)

Fig. 2 Determining the shortest route from point A to point B by physical manipulation of a mildly abstracted representation of a route network (a): The strings corresponding to route segments preserve the relative distance relations of the original route segments; the distance relations are invariant wrt. physical manipulations (pulling apart strings at A' and B') which distort angles and shapes of the route network (b) and (c). The shortest route is identified as the route corresponding to the straight connection between A' and B' in (c).

For other spatial problems, it may be easier to identify suitable spatial perspectives empirically *in the field* than analytically by computation *on a representation*. Therefore we may be better off by allowing certain operations to be carried out situation-based in physical spatial configuration as part of the overall problem solving process. With 3D printer technology, for example, we could 'print' a route network as in Fig. 2; a robot could pull apart the nodes corresponding to starting point and destination to determine the shortest route straightforwardly.

## Acknowledgements

I acknowledge discussions with Holger Schultheis, Ana-Maria Olteteanu, and the ImageSpace project team of the Collaborative Research Center SFB/TR 8 Spatial Cognition. This work is generously supported by the German Research Foundation (DFG).

## References

Braitenberg V, 1984, *Vehicles: Experiments in synthetic psychology.* Cambridge, MA: MIT Press.

Brooks RA, 1991, Intelligence without representation, *Artificial Intelligence* 47, 139-159.

Dewdney AK, 1988, *The armchair universe*, San Francisco: W.H. Freeman & Company.

Freksa C, 2013, Spatial Computing – How spatial structures replace computational effort. In Raubal M, Mark D, Frank A, eds, *Cognitive and linguistic aspects of geographic space.* Heidelberg: Springer.

Freksa C, Schultheis H, in press. Three ways of using space. In: Montello DR, Grossner KE, Janelle DG, eds, *Space in mind: Concepts for spatial education*. Cambridge, MA: MIT Press.

Marr D, 1982, *Vision,* Cambridge, MA: MIT Press.

Norman DA, 1980, *The psychology of everyday things,* New York: Basic Books, Inc.

Pfeifer R, Scheier C, 2001. *Understanding intelligence,* Cambridge, MA: MIT Press.

Searle J, 1980, Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3: 417-457.

Sloman A, 1985, Why we need many knowledge representation formalisms. In Bramer M, ed, *Research and development in expert systems,* 163-183, New York: Cambridge University Press.

# Cartographic Generalisation Aware of Multiple Representations

J.-F. Girres[1,2], G. Touya[2]

[1]UMR GRED – Université Paul Valéry Montpellier III, Route de Mende 34199 Montpellier Cedex 5
Email: firstname.name@univ-montp3.fr

[2]COGIT – IGN France, 73 avenue de Paris 94165 Saint-Mandé France
Email: firstname.name@ign.fr

## 1. Introduction

Cartographic generalisation helps deriving maps at smaller scales from a detailed geographical dataset. It is more and more frequent to have at disposal several datasets at different levels of detail in a web mapping application. For instance, a source dataset is used for deriving maps from 1:50k to 1:250k and another less detailed dataset is used to derive maps below 1:250k. Deriving intermediate scales can be helpful to generate intermediate zoom levels in a multi-scale geoportal. However, current solutions only use one dataset as input, which may lead to inconsistencies when the user switches to maps derived from a different source dataset (Figure 1).

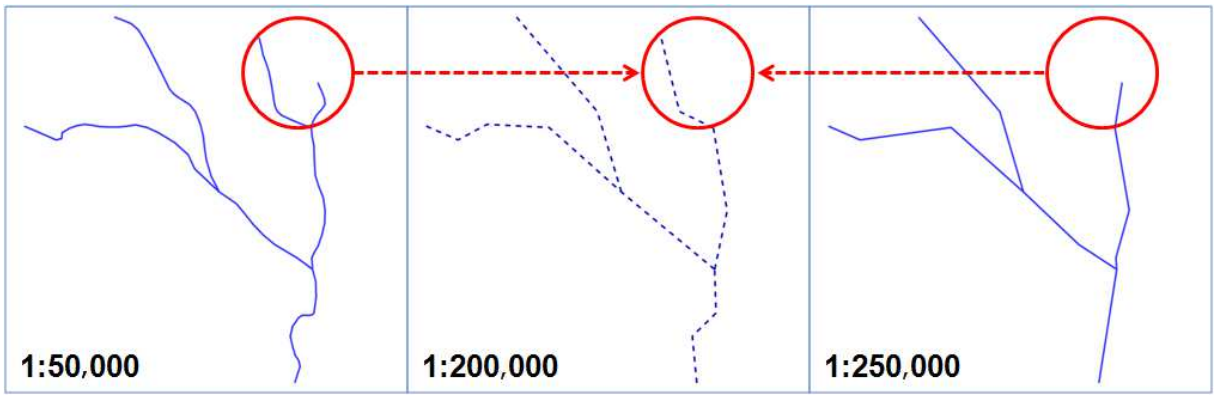



Figure 1: The intermediate scale (1:200k) is generalised from 1:50k but inconsistently with 1:250k.

The aim of the on-going research presented in this paper is to derive consistent intermediate cartographic representations to enable smooth transitions in a multi-scales geoportal. We call Multi-Representation Aware (MR-Aware) generalisation such generalisation. This work requires two major hypotheses:

- a multi-scales generalisation system to compute intermediate scales is available, like the ScaleMaster2.0 by Touya and Girres (2013), or the vario-scale model by van Oosterom et al. (2014),
- a data-matching system (e.g. Mustière and Devogèle 2006) has been used to link objects at different levels that represent the same real world entity.

The second part of the paper describes different scenarios to achieve MR-Aware generalisation. The third part describes experiments on real data and the fourth one draws some conclusions and explores further work.

# 2. Scenarios for Handling Multiple Representations during Generalisation

## 2.1 Post -processing Strategy

The first possible strategy for handling multiple representations during generalisation is to apply post-processing corrections that modify the generalised data in order to preserve consistency. There are two alternatives: a simple one and a complex one. The simple alternative is to identify the inconsistencies in the generalised output (Figure2a) and then use the next level representation to enrich the generalised output. In Figure 2, the inconsistency is a missing river that is added in the post-process.

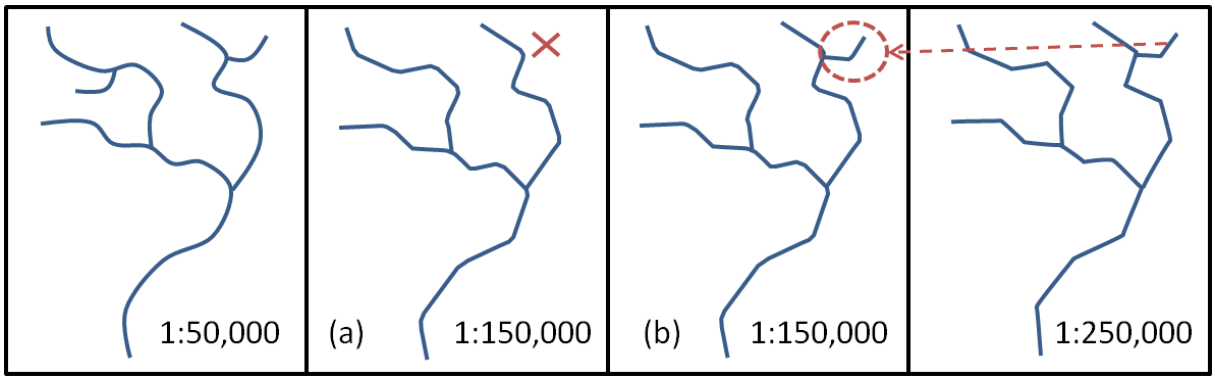



Figure 2: Post-processing strategy: the forgotten river (a) is added after generalisation at 1:150k from the 1:250k level data (b).

This scenario tends to increase the amount of data in the generalised output, which is not desirable. The second alternative is a more complex post-process that deals with this problem by reducing the amount of data after consistency has been achieved. In Figure 2, it would remove another river that is not present at the 1:250k level.

## 2.2 Pre-processing Strategy

The second strategy seeks to handle consistency between scale levels before generalisation. Once again, two alternatives are discussed. The first one consists in identifying the objects in the initial level that are linked to an object in the upper levels, and then apply generalisation only on those objects that are not linked (Figure 3): matched objects cannot be deleted as they are not processed by generalisation.

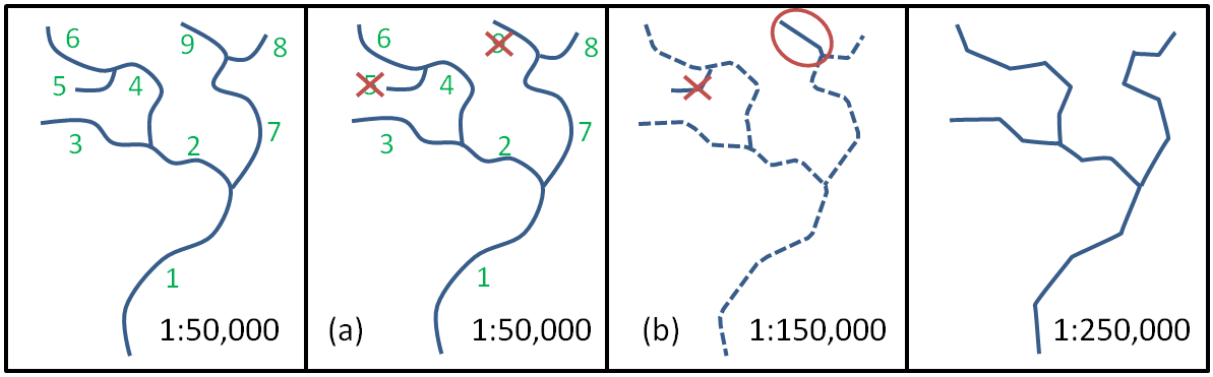



Figure 3: Pre-processing strategy: (a) the unmatched objects (5 and 9) are identified (b) selection is computed on unmatched objects only (5 is deleted and 9 is kept).

The second alternative is quite different from all other strategies, as it involves the modification of the generalisation process, while the others just provided adaptations to what a generalisation process can achieve. With this strategy, the matched objects are a complementary input of the process that has been changed to cope with a set of matched objects and a set of unmatched objects. For instance, instead of just simplifying the geometry of a matched object, the modified process will instead compute an intermediate geometry between the detailed and the undetailed matched geometries.

### 2.3 Scenarios Comparison

All four scenarios have advantages and drawbacks. They are analysed in relation to the quality of the MR-aware generalisation they can provide, and to the cost of their implementation in a given multi-scales generalisation system. The best scenario in terms of output quality is the last one that modifies the generalisation processes to take matched objects into account. Unfortunately, it is also the most costly scenario as it requires the re-implementation of each generalisation process, which is sometimes not possible, for instance in a system based on external generalisation web services (Regnauld et al. 2014). On the other hand, the simpler scenario is the post-processing addition of missing objects. We believe that the worst theoretical scenario in terms of output quality is the first pre-processing strategy where only the unmatched features are generalised. As generalisation is a holistic process, removing the neighbours of an object may lead to poor generalisation results. Finally, the most balanced scenario is the post-processing strategy that preserves consistency as well as the final amount of data in the map.
All four scenarios are sensible to the errors of the matching process, but the sensibility of each scenario to omission and commission has to be studied further.

## 3. Experiments

To illustrate MR-aware generalisation (using post-processing strategy), an experiment is provided on a sample of road networks extracted from datasets at 1:50k and1:250k (Figure 4). A preliminary matching of homologous road objects was achieved using the Mustière and Devogèle (2006) algorithm.

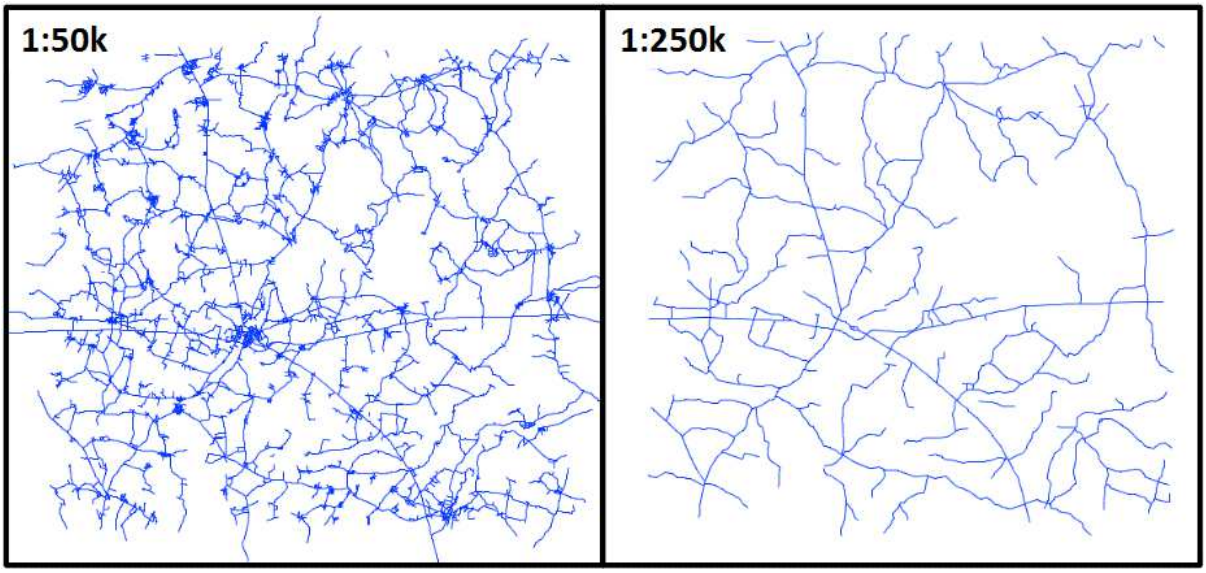



Figure 4. Road networks extracted from 1:50k and 1:250k datasets

In order to generalise an intermediary road network at the scale 1:150k, strokes-based generalisation is carried out (Thomson & Brooks 2000). The process is applied with and without MR-aware generalisation. Figure 5 shows the roads which have been preserved (in green) by MR-aware generalisation, but would have been eliminated without (in red).

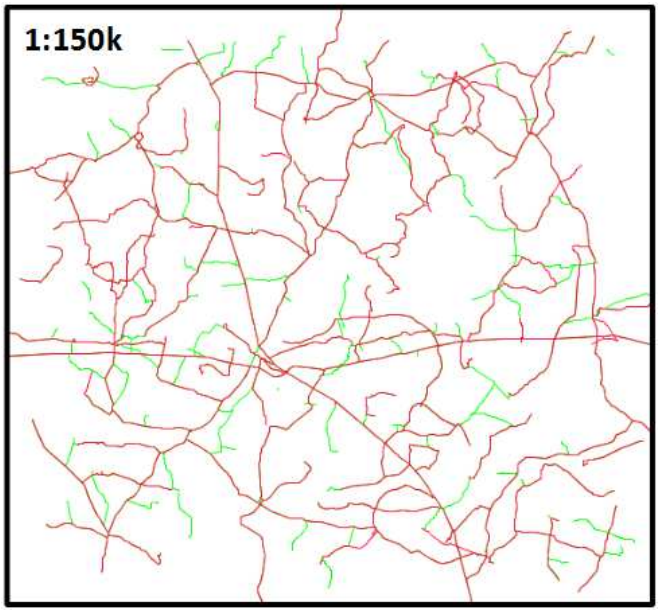



Figure 5. A road network (1:150k) without (red) and with (green) MR-aware

To quantify these differences, Table 1 exposes the difference in object numbers and roads total length, between both original datasets and the generalised road network with or without MR-aware generalisation. The results show that more than 400 km would have been deleted by not applying MR-aware generalisation.

Table 1. Comparison of generalised roads with or without MR-aware.

| Dataset | Number of objects | Length (km) |
|---|---|---|
| **Roads 1:50k** | 6512 | 5039,76 |
| **Roads 1:150k** (without MR-aware) | 2264 | 2041,64 |
| **Roads 1:150k** (with MR-aware) | 2713 | 2476,88 |
| **Roads 1:250k** | 812 | 1768,39 |

A second experiment was carried out on railroad network generalisation (Touya & Girres 2014) with a comparison of the strategy where only unmatched features are generalised and the previously tested post-processing strategy (Figure 6). Both strategies provide better results than only generalisation, and in this case, pre-processing deletes more features as removing the matched features damages the geographic context used by generalisation.

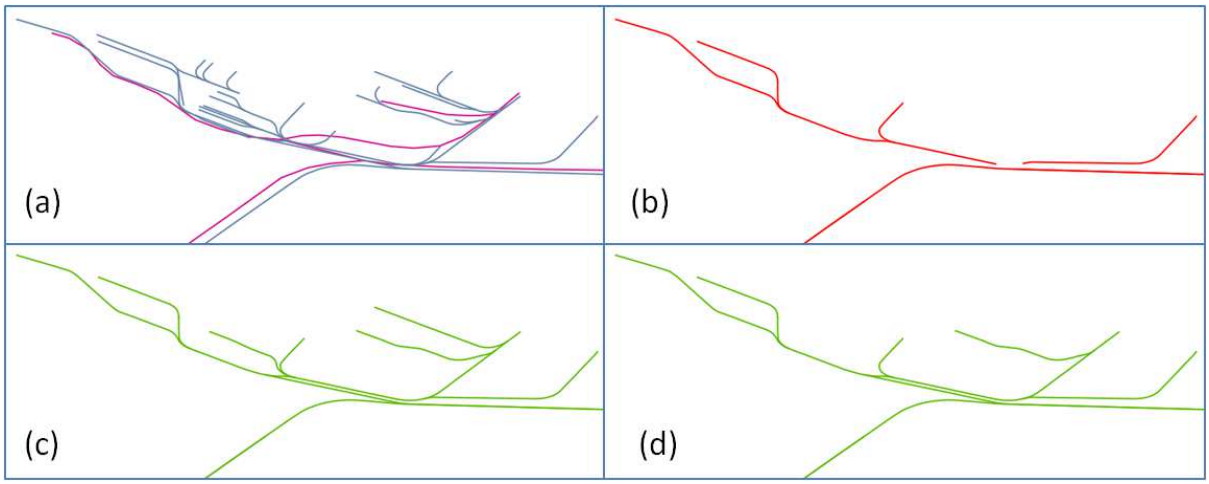



Figure 6. (a) railroad network at initial scales (1:50k in blue, 1:250k in magenta). (b) generalisation without MR-aware process. (c) post-processing MR-aware generalisation. (d) pre-processing MR-aware generalisation.

## 4. Conclusion and Further Work

This paper proposed different scenarios to enable the derivation of consistent intermediate cartographic representations between existing multi-scale levels. Two of the scenarios have been implemented and tested on real datasets, with promising results.

As the presented work is on-going research, there is much to explore. First, both implemented scenarios were tested with simple generalisation processes, and further testing should be made with more complex processes. For instance, polygon to line collapse (Figure 6) should be hard to handle with the post-processing strategy. Then, all four strategies should be tested and compared to get a clearer view on the best strategies.

Generalisation is a holistic process that requires the modelling of the geography around each object, notably the spatial relations with neighbours. Roads are drawn in a map to show they allow the access to some place, so geographical context has to be integrated to MR-aware generalisation to improve the quality of intermediate levels.

Finally, the proposed scenarios do not handle inconsistencies between levels, which occur with real datasets. In Figure 4, the bottom left road of the 1:250k dataset does not exist at 1:50k. It is not possible here to preserve consistency.

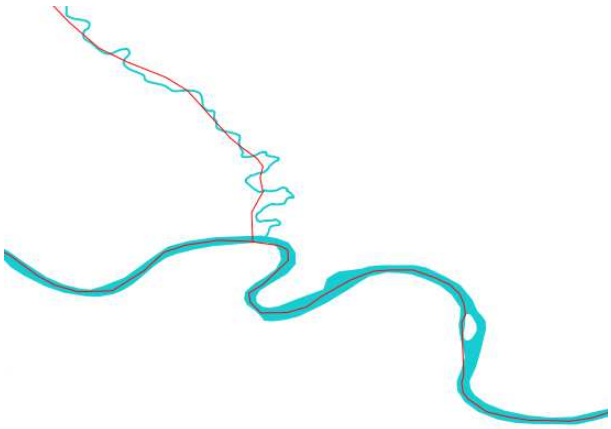


Figure 6. Rivers represented by lines and polygons at 1:50k and by only simplified lines at 1:250k.

## References

Mustière S and Devogèle T, 2006, Matching Networks with Different Levels of Detail. GeoInformatica 12(4):435–453.

Regnauld N, Touya G, Gould N and Foerster T, 2014, Process Modelling, Web Services and Geoprocessing. In: Burghardt D, Duchêne C and Mackaness W (eds), Abstracting Geographic Information in a Data Rich World. Springer, Berlin, 197–225.

Thomson R C, Brooks R, 2000. Efficient generalisation and abstraction of network data using perceptual grouping. In: Proceedings of Geocomputation. University of Greenwich, Kent, UK.

Touya G, Girres JF, 2013, ScaleMaster 2.0: a ScaleMaster Extension to Monitor Automatic Multi-Scales Generalizations. Cartography and Geographic Information Science 40(3):192–200.

Touya G, Girres JF, 2014, Generalising Unusual Map Themes from OpenStreetMap. Submitted to the 17th ICA Workshop on Generalisation and Multiple Representation.

van Oosterom P, Meijers M, Stoter J and Suba R, 2014, Data Structures for Continuous Generalisation: tGAP and SSC. In: Burghardt D, Duchêne C and Mackaness W (eds), Abstracting Geographic Information in a Data Rich World. Springer, Berlin, 83–117.

# Feature Selection for Land Use Classification Based on Temporal Activity Patterns

L. Gong, X. Liu, Y. Liu

Institution of Remote Sensing and Geographical Information System, Peking University, Beijing 100871
Email: {gongli.mxj; 1989liuxi; liuyu}@{gmail.com; gmail.com; urban.pku.edu.cn}

## 1. Introduction

In the era of big data, large scale geo-tagged data, such as taxi trajectory data, mobile phone communication data and check-in record data in social network, are widely used to study urban problems in the perspective of urban planning, traffic management and so on. A lot of literature (Liu et al. 2012, Pei et al. 2013) applies these big geo data to land use classification, as they find that temporal fluctuations of residents' activity intensity are different between different land use types, but similar among the same type. Usually the temporal changes of activity volume are constructed as a feature vector for each place. Then supervised or unsupervised algorithms are adopted to infer the land uses. However, there are few attempts to explore the feature selection problem. Intuitively we think that not all of those features contribute to land use detection. Choosing some crucial ones would reduce the running time of learning algorithms and lead to better understanding of the relationship between citizens' activities and social function of places. In this study, we apply the ReliefF algorithm to estimate the importance of features and select key feature subsets. Previous research is normally based on one kind of activity data. We think different activity data also present various temporal changes, so we conduct experiments on three kinds of activity datasets in order to generate comprehensive suggestions for future research.

## 2. Data and Methods

### 2.1 Data preparation

Our research region is Shanghai, China, which is discretized into 250 $m^2$ grids. In this paper we only investigate three typical land use types: business area, commercial area and residential area. Related points of interest (POIs) are aggregated into three groups to determine the land use of each grid (Table 1). We adopt the method in Reades et al. (2009) to measure the relative concentration of each kind of POIs. The dominant POIs determine the land use types of grids. We pick 100 grids from each land use type as the sample dataset $G = \{g_i | i = 1,2,\dots,300\}$ under the guidance of planning maps and cognition of research region.

In this study three kinds of activities are organized, including pick-up and drop-off events extracted from taxi trip data during one week, and check-in records over one year. These activity data have common attributions on time and location. For a given grid $g_i \in G$, the average activity volumes of pick-up, drop-off and check-in activity at time $t$ are defined as $p_i^t$, $d_i^t$ and $c_i^t$ respectively. Next, we construct three feature vectors for $g_i$: $P_i = [p_i^1, \dots, p_i^{24}]$, $D_i = [d_i^1, \dots, d_i^{24}]$ and $C_i = [c_i^1, \dots, c_i^{24}]$. Normalization is needed to eliminate the differences in magnitude among grids and keep the shape of temporal signature. Taking $P_i$ as an example, the equation is ${p_{norm}}_i^t = (p_i^t - \mu_i)/\sigma_i$, where $\mu_i = \sum_{t=1}^{24} p_i^t/24$, $\sigma_i = \sqrt{\sum_{t=1}^{24}(p_i^t - \mu_i)^2/23}$.

Finally, we get three datasets based on pick-up, drop-off and check-in activities which are defined as $S_p = \{P_i^{norm}\}$, $S_d = \{D_i^{norm}\}$ and $S_c = \{C_i^{norm}\}$ respectively. There are 300 samples with predetermined land use types in each dataset.

Table 1. The representative POIs of land use types

| Land Use | POI |
|---|---|
| Business Area | Office Building; Government Office; Company; |
| Commercial Area | Shopping Mall; Market; Restaurant; |
| Residential Area | Residential District; |

### 2.2 ReliefF algorithm

Relief algorithm is a general and successful attribute estimator, but limited to work only for binary classification. ReliefF algorithm is an extension of Relief, which can deal with multi-class situation. The basic idea of ReliefF is to estimate the weight of attributes according to how well their values distinguish between instances that are near to each other (Dash and Liu 1997). For that purpose, Relief randomly chooses a sample $g_i$ from the training dataset. Then it searches for $k$ nearest neighbours of the same class, named Nearest Hits $H_i$, and $k$ nearest neighbours from each of other classes, called Nearest Misses $M_i$. The weight of attribute $A_j$ is increased if the values of $g_i$ and $H_i$ are more similar on $A_j$, and reduced if the values of $g_i$ and $M_i$ are more similar. The whole process is repeated for $m$ times. After that, we choose all features having weight larger than or equal to a threshold $\theta$, where $m$ and $\theta$ are user-defined parameters.

## 3. Experiments and Results

Our experiments were conducted on three kinds of activity datasets. For each dataset, the same processing chain was followed: 1) evaluating and selecting the features that distinguished one specific land use type well; 2) evaluating and selecting the feature subset that discriminated all classes well. In the first step, we chose one land use type from business, commercial and residential as the target object, and merged samples of the other two classes. Then we got the feature estimation result that descripted features' capacity to represent the target class. The higher the weight is, the more important the feature is. In the second step, the key features that contributed to distinguishing three land uses were explored. To validate the selected feature subset, k-nearest neighbour classifier was adopted to test the classification accuracy of feature subset. Unless otherwise stated, the same learning settings were used for all datasets, namely: $k_R = 10$, $k_N = 8$, which are the number of k-nearest instances in Relieff and KNN respectively.

**Pick-up activity**

Figure 1 shows the weight of features for different classification purposes based on pick-up activity. The feature in 8 a.m. is important for all land uses. In addition, the feature in 9 o'clock in the morning outstands for residential area, because there are many people take a taxi to work at this time. The representative feature for business region is 4 p.m. that corresponds to people leaving workplaces. Commercial area's key feature is 9 o'clock in the evening when citizens would go back home. We calculate the number of instances that are classified into wrong class based on first k important features. As can be seen from Figure 2, first ten key features can replace all the features, as the false rates keep steady from $k = 10$.

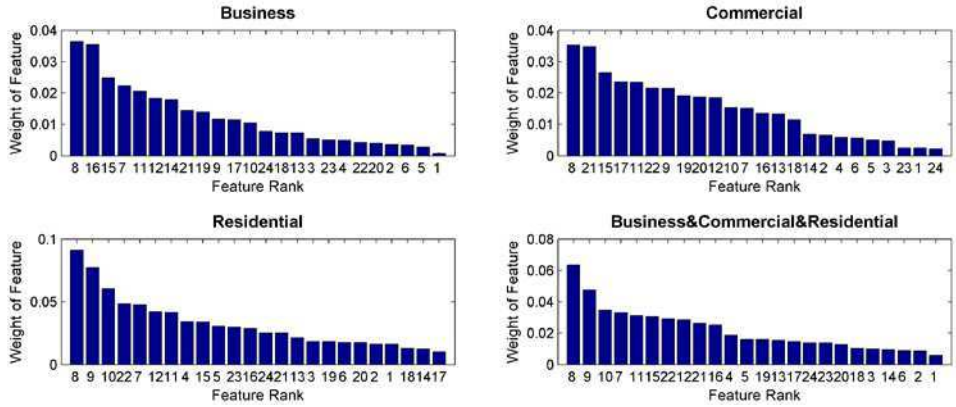



Figure 1: The weight of features for different classification purposes based on pick-up activity

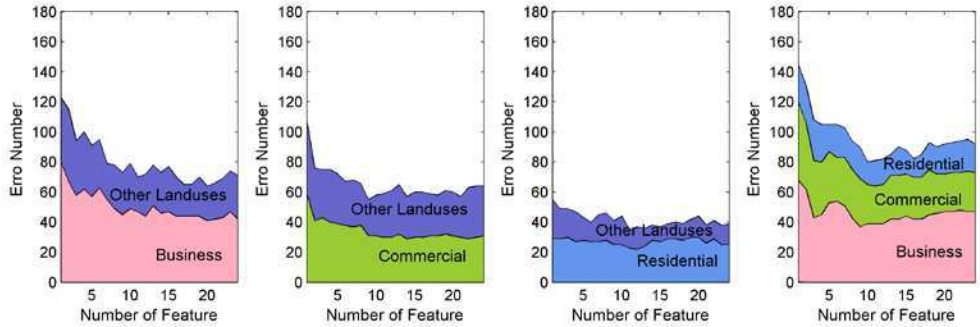



Figure 2: The variance of false rate with increasing the number of key features for different classification purposes based on pick-up activity

**Drop-off activity**

For drop-off activity, the key features of three land use types are different. Business area shows different pattern from other land uses from 9 a.m. to 10 a.m. The high weight features are found at noon and evening time in commercial area. We should note that although we can gain more relevant features to detect residential instances, the validity result indicates that these feature subsets can not explain most information of residential land use well (see in Figure 4).

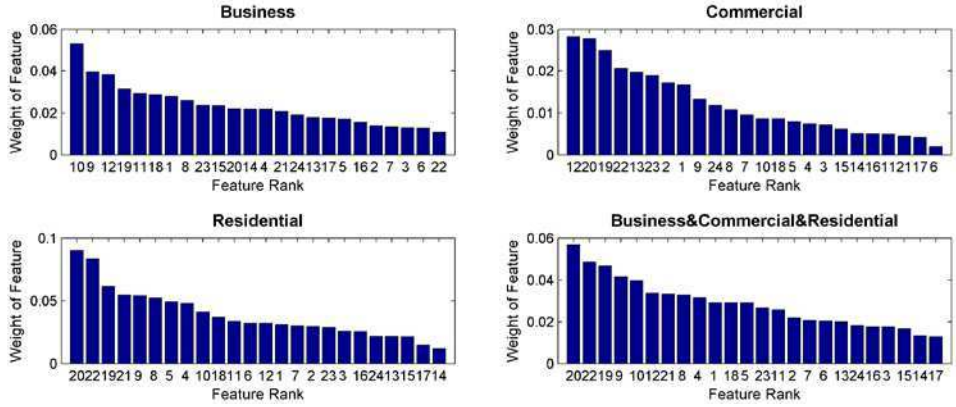



Figure 3: The weight of features for different classification purposes based on drop-off activity

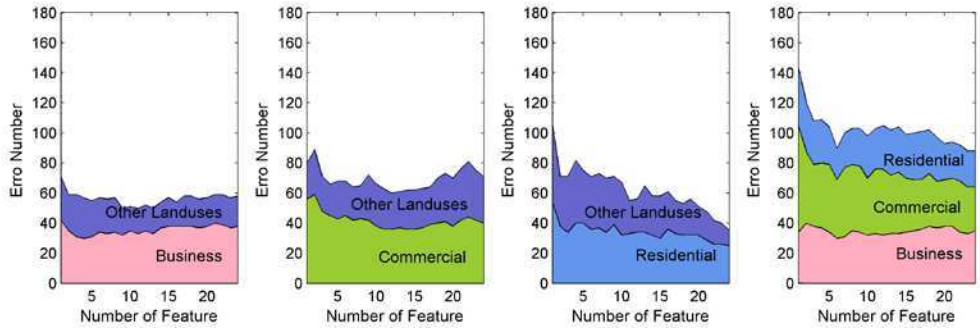



Figure 4: The variance of false rate with increasing the number of features for different classification purposes based on drop-off activity

**Check-in activity**

The activity volume in 7 p.m. is most relevant to classify land uses. Except that, the key features of business area mainly distribute in daytime, while the other two types act significant differences at night. Figure 6 shows that it's necessary to make feature selection on check-in dataset, as classifier accuracies are similar between key feature subsets (about or more than 5 key features) and the complete feature set.

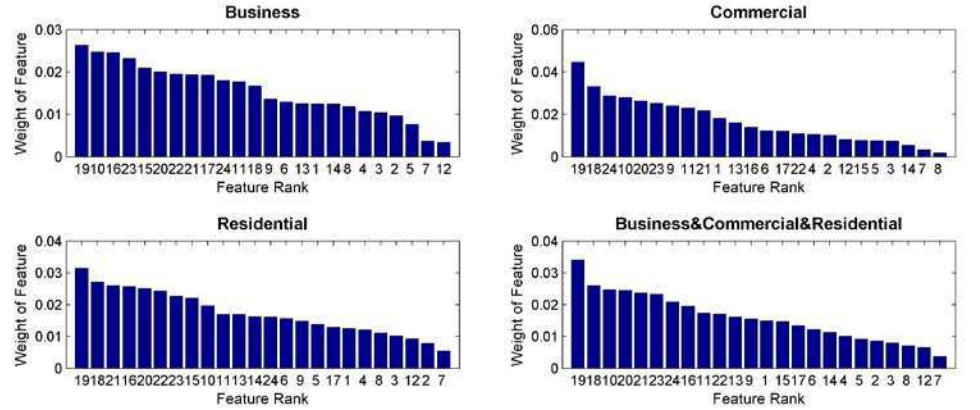



Figure 5: The weight of features for different classification purposes based on check-in activity

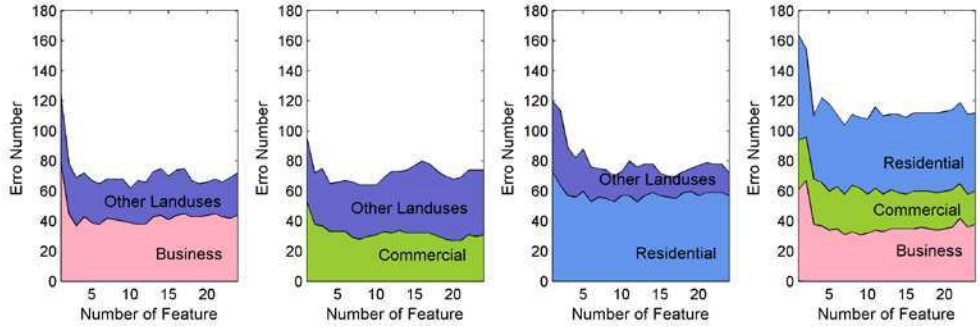



Figure 6: The variance of false rate with increasing the number of features for different classification purposes based on check-in activity

## 4. Conclusion

This work explored the relationship between temporal fluctuations of activity intensity and land use types. We applied feature selection methods to extract the representative feature subsets that can distinguish land uses well. As different activities had different fluctuations in hourly amount for one land use type, a comparative experiment was conducted on pick-up,

drop-off and check-in datasets. We found out some key features of each land use type. The experiment results also indicated that feature selection processing was needed to be considered when using temporal changes of human activity intensities to infer land use types.

## References

Dash M and Liu H, 1997, Feature selection for classification. *Intelligent data analysis*, 1(3): 131-156.

Liu Y, Wang F, Xiao Y and Gao S, 2012, Urban land uses and traffic 'source-sink areas': evidence from GPSenabled taxi data in Shanghai. *Landscape and Urban Planning*, 106: 73-87.

Pei T, Sobolevsky S, Ratti C, Shaw SL and Zhou C, 2013, A new insight into land use classification based on aggregated mobile phone data. *arXiv preprint arXiv*:1310.6129.

Reades J, Calabrese F and Ratti C, 2009, Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5): 824-836.

# Generating Large-scale and Health-related Synthetic Population Microdata at a Neighbourhood Level in Japan

K. Hanaoka[1], T. Nakaya[2], T. Tabuchi[3]

[1]International Research Institute of Disaster Science, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, Japan
Email: hanaoka@irides.tohoku.ac.jp

[2]Department of Geography, Ritsumeikan University, 58 Komatsubara Kitamachi, Kita-ku, Kyoto, Japan
Email: nakaya@lt.ritsumei.ac.jp

[3]Center for Cancer Control and Statistics, Osaka Medical Center for Cancer and Cardiovascular Diseases,
1-3-3 Nakamichi, Higashinari-ku, Osaka, Japan
Email: tabuchitak@gmail.com

## 1. Introduction

The purpose of this study is to construct large-scale and health-related synthetic population microdata using a spatial microsimulation approach for understanding detailed geographic variations in health-related status and behaviour, such as smoking and health examinations, at the neighbourhood level in Japan.

The spatial microsimulation method creates geographically disaggregated microdata by combining multiple data sources. Morrissey (2008) and Smith (2009) developed spatial microsimulation models for health planning in small-sized study regions. Unlike previous studies, a challenging task of our study is to construct microdata for the entire Japanese population with living neighbourhood information, constituting microdata for over 128 million people from 208,476 neighbourhoods throughout Japan. The data structure and algorithm of the spatial microsimulation method were reviewed to minimize computational time and make it possible to create such large-scale synthetic population microdata.

The neighbourhood has recently gained status in the geography, sociology, economics, and public health fields as a key term for understanding health status, poverty, fear of crime, and education (van Ham et al. 2012). Due to the similarities in lifestyle and health services in neighbourhoods, the local population often share similar health behaviours and problems. From a policy standpoint, understanding the geography of relevant health measures and the relative position of a concerned neighbourhood is therefore important for spatially matching health care needs, supplies, and provisions.

It remains difficult, however, to obtain such large-scale microdata with health-related attributes because (1) the microdata from a population census is often unavailable to academic researchers, as well as local government officers, (2) sampling surveys do not have an adequate number of samples from all neighbourhoods due to budget constraints and limited available human resources, and (3) often these surveys contain various sampling biases (e.g., the elderly in rural areas are more likely to respond to questionnaires than the young in urban locales), which results in skewed output indices. Therefore, to overcome these issues, in this article, we employed a spatial microsimulation model to construct large-scale and health-related synthetic population microdata for the entire Japanese population.

# 2. Method

## 2.1 Procedure

Figure 1 illustrates the procedure for creating a set of synthetic population microdata in this study. Spatial microsimulation is defined as a method to find the best combination of survey samples (seed) so that the aggregated tables agree with the benchmark small-area census tables (benchmark). We used Simulated Annealing (SA), a combinatorial optimisation algorithm, to collect samples from the seed survey samples repeatedly until the aggregates of the synthetic population microdata agree with the benchmark tables. This procedure is equivalent to updating the weights of survey samples, but it can adjust sampling biases multidimensionally by considering multiple benchmark tables.

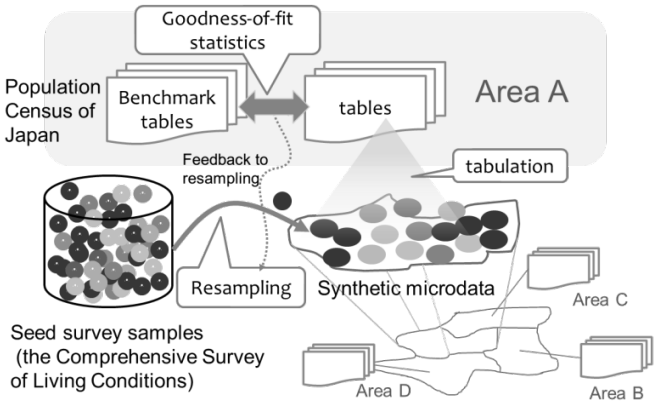



Figure 1. Illustrative explanation of synthetic microdata generation in spatial microsimulation

## 2.2 Datasets and Data Structure

For the seed survey samples, we included approximately 270,000 household samples from the Comprehensive Survey of Living Conditions 2010 in the model. This survey is the largest sampling survey of individual health conducted by Japan's Ministry of Health, Labour and Welfare. It contains the same household/individual variables as those of a population census, as well as variables on health status. This dataset was divided by prefecture and the prefectural samples were used as the seed survey samples for each prefecture.

As the benchmark tables, we selected five tabulations of persons and three tabulations of households from the 2010 Population Census of Japan. All tabulations were compiled at "Cho-cho-aza", a neighbourhood area of the Japanese census. The selected benchmark tables at the person-level were (1) sex by age, (2) sex by occupation, (3) sex by marital status, (4) sex by education, and (5) sex by employment, and, at household-level, (6) housing tenure, (7) dwelling type, and (8) family type.

A bottleneck of the SA algorithm is its long computational time. In microsimulation, the microdata of persons and households are often stored in a list-based format (one row represents one person or one household). Since both person- and household-level variables need to be included, all person records have to be updated and then re-tabulated to compute goodness-of-fit statistics if a new household is selected in SA. Thus, to avoid unnecessary tabulations during SA iterations, we calculated the frequencies of person records beforehand and combined these person-level variables with a household record. The structure of this data is shown in Figure 2.

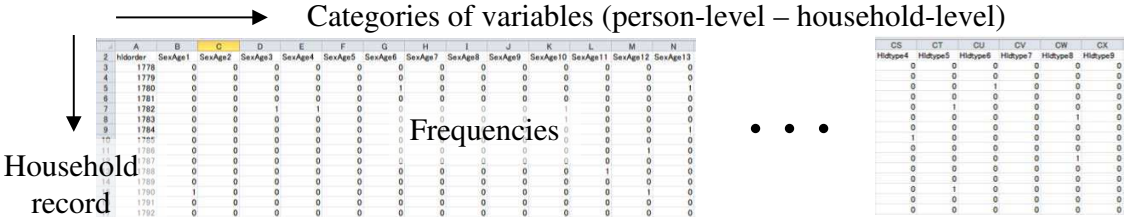



Figure 2. Data structure (seed survey samples)

## 2.3 Generation of Synthetic Population Microdata

The SA algorithm was applied for each neighbourhood. The parameters in SA were set as follows: number of maximum iterations: 8 million times; initial temperature: 109; and temperature reduction ratio: 0.05. To reduce computation loads, goodness-of-fit statistics during the SA iterations were based on the Overall Total Squared Error (Overall TSE) rather than the Relative Sum of Squared Z-score (RSSZ). The overall TSE is defined as follows (Equations 1 & 2):

$$\text{TSE}^k = \sum_{ij} (E_{ij}^k - O_{ij}^k)^2 \quad (1)$$

$$\text{Overall TSE} = \sum_k TSE^k \quad (2)$$

*k*: *k*th benchmark table, *i*: *i*th column, *j*: *j*th row, E: estimated counts, O: observed counts

According to Williamson (2012), other goodness-of-fit statistics, such as the Total Absolute Error (TAE), Overall TAE, TAE per household, RSSZ, Overall RSSZ, and Non-Fitting Tables (NFT), were computed after synthetic microdata were completed for a neighbourhood. In particular, RSSZ is based on Chi-square statistics and, if RSSZ exceeds 1 (i.e. exceeds the 5% $\chi^2$ critical value), the table tabulated from the synthetic microdata and its corresponding benchmark table are deemed not to be the same.

By using a standard workstation with two Xeon 6 core 2.66GHz CPUs, 48GB RAM memory, and 20 threads, the total computational time was 605,189 seconds, which is equivalent to an average of about seven days and 60 seconds per neighbourhood. Therefore, considering the size of the population and neighbourhoods to be estimated and the number of benchmark tables, we considered the total computational time to be within an acceptable range, although we will need to conduct a detailed examination of computation time by procedure.

The results of the overall goodness-of-fit statistics are presented in Table 1. The average of overall TAE across whole neighbourhoods was 31.5 counts and OTAE (sum of TAE for eight benchmark tables) per household was less than 1, which is an indication of good fit. Table 2 shows the goodness-of-fit statistics by benchmark table. The averages of RSSZ are well below 1. Evaluations based on TSE resulted in low RSSZ scores.

Table 1. Overall goodness-of-fit

| Statistics | Average |
|---|---|
| Overall TAE | 31.5 |
| OTAE per household | 0.457 |
| Overall RSSZ | 2.631 |
| Non-fitting tables | 0.086 |

Table 2. Goodness-of-fit by benchmark table

| Benchmark tables | Average TAE | Average RSSZ |
|---|---|---|
| sex by age | 11.0 | 0.228 |
| sex by occupation | 6.2 | 0.396 |
| sex by marital status | 1.8 | 0.055 |
| sex by employment | 2.6 | 0.137 |
| sex by education | 2.8 | 0.391 |
| housing tenure | 2.4 | 0.817 |
| dwelling type | 1.3 | 0.511 |
| family type | 3.4 | 0.097 |

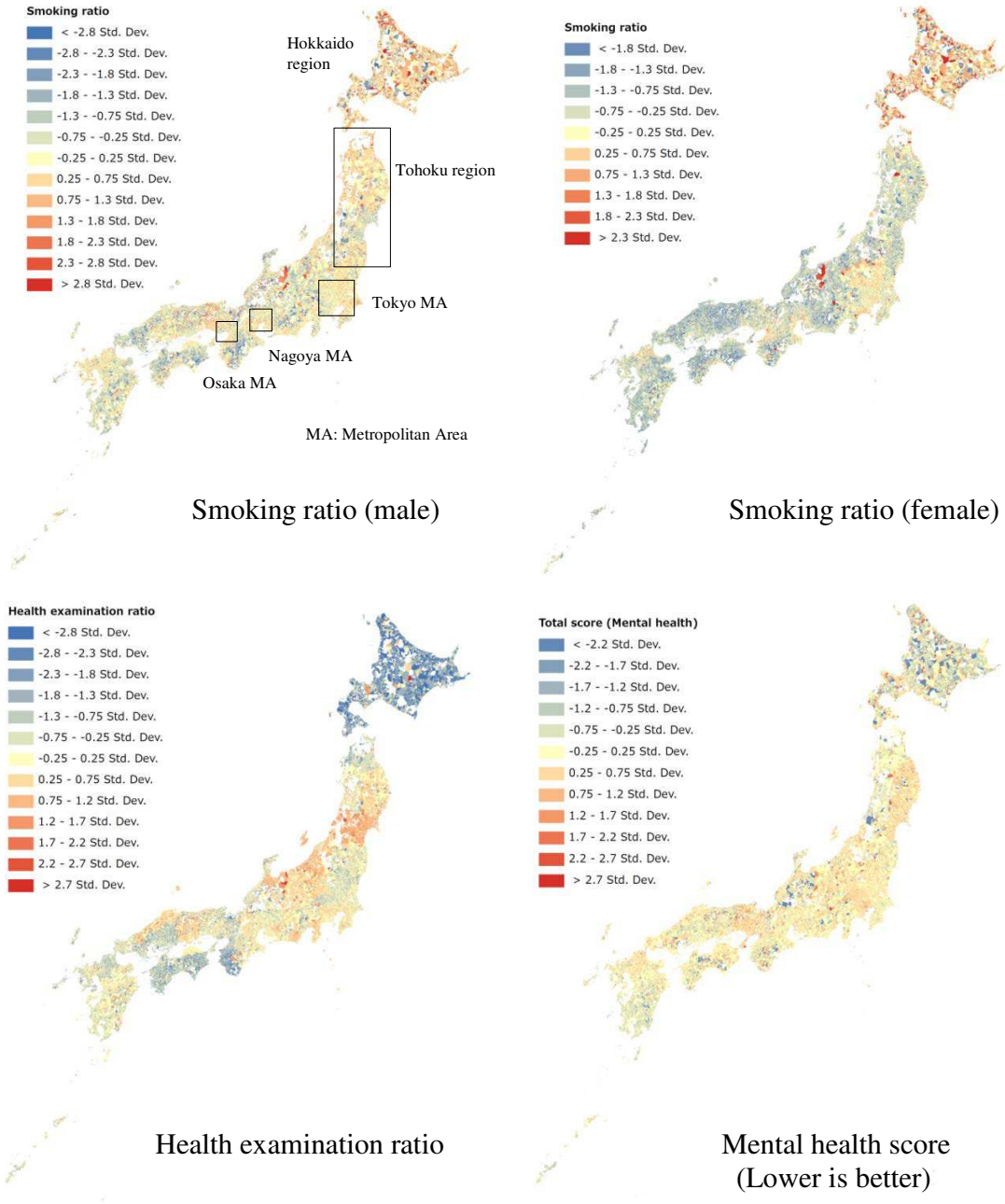



Figure 3. Geographic distributions of health-related indices at the neighbourhood level

## 3. Mapping Health-related Indices from Synthetic Microdata

We tabulated the synthetic population microdata by neighbourhood and health-related variables to map those distributions in Figure 3. First, the gender difference is obvious for the smoking ratio of people aged 20 years old and over; the smoking ratio for males is higher in both urban and rural areas, but smoking ratio among females is highly concentrated in metropolitan areas (except for the Hokkaido region). The ratio of people who had a health examination within the

past year shows discrepancies at the prefectural and neighbourhood levels, possibly due to differences in the age composition at the neighbourhood level and the effects of promotions from the prefectural government. Concentrations of neighbourhoods with a higher mental health score (K6 total score) are found in metropolitan areas and the east coast of the Tohoku region.

## 4. Conclusions

Our key findings are summarized as follows:

- Spatial microsimulation is able to combine information from survey samples and population censuses to produce highly accurate health-related synthetic population microdata. Notably, this study reviewed the data structure and algorithm to create large-scale synthetic population microdata for over 120 million Japanese people within an acceptable computation time.
- The merits of synthetic population microdata are that sampling biases are adjusted multidimensionally during the SA procedure and an analyst can tabulate them flexibly according to particular analytical needs. In this study, the spatial distributions of health-related status and behaviour were mapped at the neighbourhood level all over the country. It is now possible to examine differences in health status between urban and rural neighbourhoods or those among urban neighbourhoods in a large city. Neighbourhood statistics of health status and behaviour, and the relative position of a neighbourhood among neighbourhoods are useful for determining appropriate amounts of health services and provisions and efficient spatial allocation by local government.

## Acknowledgements

This work was supported by JSPS KAKENHI (Grant Number 24300323) and the Ministry of Health, Labour and Welfare (Grant; Comprehensive Research on Life-Style Related Diseases including Cardiovascular Diseases and Diabetes Mellitus (H25-010)). Data were used with permission from the Japanese Ministry of Health, Labour and Welfare. The analyses of national survey data were considered to be exempt from the need for ethical review according to the Epidemiological Research Guidelines.

## References

Morrissey K, Clarke G P, Ballas D, Hynes S and O'Donoghue C, 2008, Examining access to GP services in rural Ireland using microsimulation analysis. *Area*, 40(3): 354-364.

Smith D M, Clarke G P and Harland K, 2009, Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A*, 41(5): 1251–1268.

van Ham M, Manley D, Bailey N, Simpson L and Maclennan D, 2012, Neighbourhood effects research: New perspectives. In: van Ham M, Manley D, Bailey N, Simpson L and Maclennan D (eds), *Neighbourhood Effects Research: New Perspectives*, Springer, New York, USA, 1-21.

Williamson P, 2012, An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation. In: Tanton R and Edwards K L (eds), *Spatial Microsimulation: A Reference Guide for Users*, Springer, New York, USA, 19–47.

# An Ontological Solution for Perceptual Uncertainties of VGI

S. Hassany Pazoky, F. Karimipour, F. Hakimpour

Department of Surveying and Geomatics Engineering, University of Tehran, Tehran, Iran
{shpazooky; fkarimipr; fhakimpour}@ut.ac.ir

## 1. Introduction

GIScience is experiencing a thorough revolution influenced by the advances in digital era (Sui et al. 2013). In particular, data acquisition, which has always been the costly part of geospatial projects, has benefited the most (Goodchild 2009). As an emerging instance, Volunteered Geographic Information (VGI), coined by Goodchild (2007a), conceptualizes the use of unbeatable power, knowledge, expertise, and ubiquity of the general public in spatial data acquisition.

VGI leads to providing a huge amount of data in a cheap manner, without interfering in the acquisition process. However, it is still questionable in terms of quality, heterogeneity, and integration (Goodchild 2008, Goodchild and Li 2012, Roche et al. 2012, Elwood et al. 2013, Fairbairn and Al-Bakri 2013, Feick and Roche 2013, Haklay 2013, Hodgson et al. 2014). In particular, this paper introduces human perception as a significant source of VGI uncertainties. As VGI relies on human as sensors (Goodchild 2007a, 2007b), different perceptions of the same facts make VGI imbalanced. We propose an ontological solution for this issue. The problem is discussed, and the solution is described through a simple example.

## 2. Perceptual Uncertainties of VGI

Many scholars have stated their concern on VGI uncertainties, but only few practical solutions have been recommended. All these concerns rise due to the fact that VGI is acquired by the public "with no brand, no experience or training, and no standards" (Goodchild 2008). Data acquired by experts always follows pre-defined standards and are accompanied by metadata (Goodchild 2009), which is absent in VGI. Participation of a diverse crowd from different walks of life in VGI production is troublesome from an uncertainty point of view (Roche et al. 2012). Although several papers have been devoted to quality assurance and providing metadata for VGI (Goodchild 2007a, 2007b, 2008, Goodchild and Li 2012, Elwood et al. 2013, Karimipour et al. 2013, Sui 2014), it is still an open issue and no conclusive framework is widely accepted. As a specific case, this paper focuses on VGI uncertainties caused by diverse (linguistic) perceptions of public participants, and presents the initial results of an ontological solution for this issue.

A verbal communications is simply considered as: 1) Expressing the idea in words by the first party, 2) Translating from the target to the destination language; and 3) Comprehending the translated words by the second party. In the case of VGI platforms, different types of uncertainty may occur in the three steps:

- The labels provided to be assigned to the objects (e.g., highway, motorway, etc.) may differ from what the designer really mean. For instance, is "highway" the exact term for what is intended?
- In case of multilingual VGI portals, translating the labels to the destination languages may deviate from its original concept. Even, the words used in countries with similar language may differ, e.g. motorway (England), motorways, freeways, and freeway-

like roads (Australia), limited-access highway (Canada), freeway (India), and Limited access freeway (USA)[1].

- As there are no pre-defined standards, the labels are chosen by the users based on their own perception, which certainly differs from a user to another. For example, "residential road", "primary road", "secondary road", and "tertiary road" are among the different types of linear features to be chosen in OpenStreetMap, which confuse the users. Although documentations are provided, many users do not bother themselves reading the descriptions, and simply choose the one that is closer to what they perceive. Even if they are read, users may have different interpretations of the same definition, which is regarded as ambiguity (Shi 2010).
- Different people may instantiate a certain object to different classes, which results in vagueness (Shi 2010). Again, definitions of OpenStreetMap for the road types do not share a clear boundary.
- Finally, even assuming clear definitions of object types and perfect instantiation, lack of information prevents the right class to be chosen, called impreciseness by Shi (2010). A very illustrative example is the location of Eifel tower: Although answers such as Europe, France, and Paris are all correct with no uncertainty, Europe is the most and Paris is the least imprecise notion.

Note that the above types of uncertainty only concern the non-spatial parts of geographic data. Recent progresses in automatic and web cartography have alleviated the uncertainty of drawing the spatial part (Pazoky and Hakimpour 2014). The user and the system designer exactly know what each other mean by every drawing. However, it must not be confused with positional accuracy, which is about the uncertainty in the position of the objects.

## 3. How Can Ontology Help?

To solve the issues discussed, we propose a taxonomical design, instead of highly-complex individual words. Opposed to what is currently available on, say, OpenStreetMap, the label is chosen from a multilayer hierarchy. The user is asked a simple question at each node, whose answer leads him/her to the next level; until the leaf is reached, or he/she lacks information, i.e. cannot answer the question. The labels of nodes of the hierarchy are not shown to the user, as it again causes uncertainty and may bias him/her towards some instances.

Figure 1 illustrates the proposed ontological labeling resulted in the hierarchy depicted in Figure 2. The questions are quite simple. The words used in the questions are not used in professional contexts. On the other hand, significant distinction between the choices leaves no room for vagueness. In other words, the bounds of the choices are clearly distinguishable. Finally, the users go through the hierarchy as much as they have information. Therefore, the more the feature is distant from the root, the more the user is familiar with the feature; i.e., the less the impreciseness.

## 4. Summary

VGI is one the most significant advances in GIScience following the well-known elements of Web 2.0. Although VGI is proved to be very useful in numerous applications, its uncertainty issues must be deeply investigated, as it is acquired by a diverse public. This paper opens a discussion on perceptual uncertainties of VGI and proposed an ontological solution. However, its presentation and implementation to be comprehended by ordinary people, as the main users of VGI portals, is still questionable.

[1] http://wiki.openstreetmap.org/wiki/Highway:International_equivalence

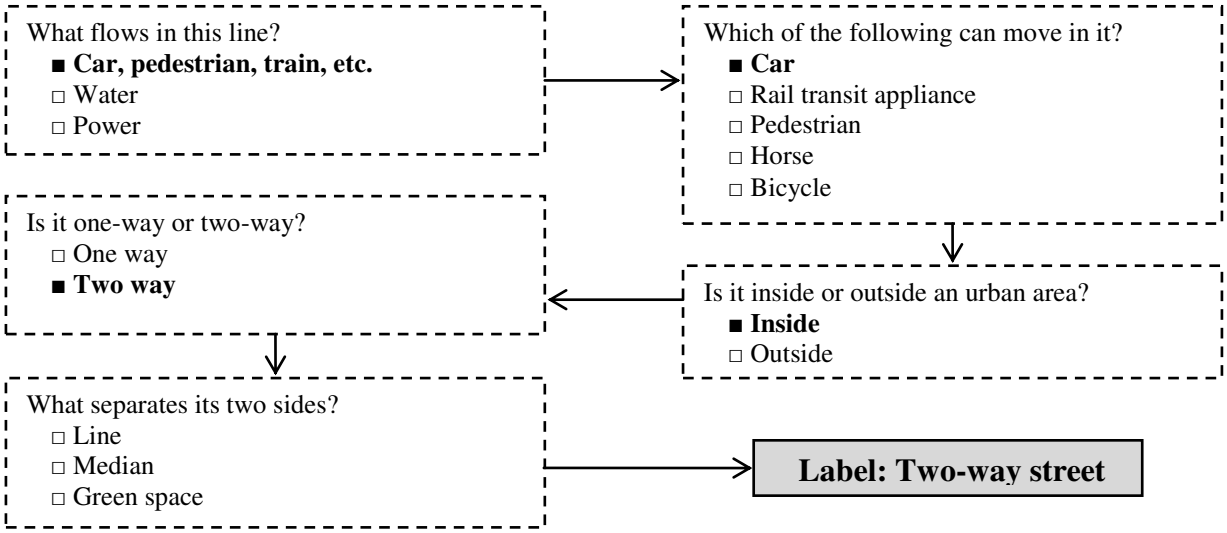



Figure 1. The proposed ontological labeling for a line (The user's answers are thickened)

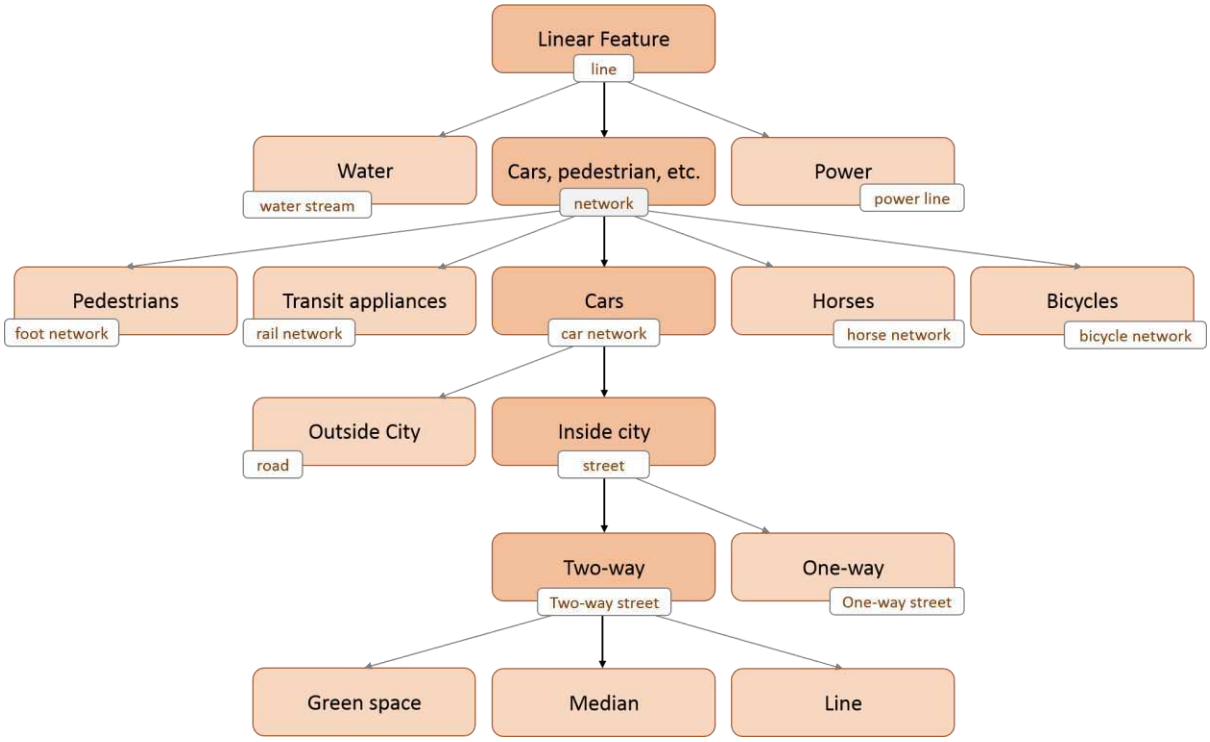



Figure 2. The hierarchy resulted by "questions and answers" shown in Figure 1. The path towards the result is thickened. Note that, the names are not shown to the user.

## References

Brown G, and Kyttä M, 2014, Key issues and research priorities for public participation GIS (PPGIS): a synthesis based on empirical research. Applied Geography, 46(1): 122-136.

Elwood S., Goodchild M F and Sui D, 2013, Prospects of VGI research and the emerging fourth paradigm. In: Crowdsourcing Geographic Knowledge, Springer Netherlands, 361-375.

Fairbairn D. and Al-Bakri M, 2013, Using geometric properties to evaluate possible integration of authoritative and volunteered geographic information. ISPRS International Journal of Geo-Information, 2(2), 349-370.

Feick R and Roche S, 2013, Understanding the value of VGI. In: Crowdsourcing Geographic Knowledge, Springer Netherlands, 15-29.

Goodchild M F, 2007a, Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), 211-221.

Goodchild M F, 2007b Citizens as voluntary sensors: spatial data infrastructure in the world Web 2.0, International Journal of Spatial Data Infrastructure Research, 2, 24-32.

Goodchild M F, 2008, Spatial accuracy 2.0. In: Proceedings of the Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 1, 1-7.

Goodchild M F, 2009, Geographic information systems and science: today and tomorrow. Annals of GIS, 15(1), 3-9.

Goodchild M F and Li L, 2012, Assuring the quality of volunteered geographic information, Spatial statistics, 1(1), 110-120.

Haklay M, 2013, Citizen science and volunteered geographic information: overview and typology of participation. In: Crowdsourcing Geographic Knowledge, Springer Netherlands, 105-122.

Harvey F, 2013, To volunteer or to contribute locational information? Towards truth in labeling for crowdsourced geographic information. In: Crowdsourcing Geographic Knowledge, Springer Netherlands, 31-42.

Pazoky SH and Hakimpour F, 2014, Transforming GML to presentation languages by extending XSLT. Journal of Geographic Information System, 6(1), 59-69.

Hodgson M E, Battersby S E, Davis B A, Liu S, and Sulewski L, 2014, Geospatial data collection/use in disaster response: A United States nationwide survey of state agencies. In: Cartography from Pole to Pole, Springer Berlin Heidelberg, 407-419.

Karimipour F, Esmaeili R and Navratil G, 2013, Cartographic representation of spatial data quality parameters in volunteered geographic information, Digital Proceedings of the 26th International Cartographic Conference (ICC 2013), Dresden, Germany.

Roche S, Mericskay B, Batita W, Bach M and Rondeau M, 2012, WikiGIS basic concepts: Web 2.0 for geospatial collaboration. Future Internet, 4(1), 265-284.

Shi W, 2010, Principles of modeling uncertainties in spatial data and spatial analyses. CRC Press.

Sui D, Goodchild M and Elwood S, 2013, Volunteered geographic information, the exaflood, and the growing digital divide. In: Crowdsourcing Geographic Knowledge, Springer Netherlands, 1-12.

Sui D, 2014, Opportunities and impediments for open GIS. Transactions in GIS, 18(1), 1-24.

# Multi-criteria aggregation for sensitive parcel-based census data

S. Indermühle[1], P. Laube[1], M. Geilhausen[1], T. Zwicker[2 3]

[1]Zurich University of Applied Sciences ZHAW, Grüental, 8820 Wädenswil, Switzerland
Email: {indu; patrick.laube; martin.geilhausen}@zhaw.ch

[2]tsquare gmbh, Arbergstr. 1, 8405 Winterthur, Switzerland
Email: tzwicker@tsquare.ch

[3]City council Männedorf, Bahnhofstrasse 10, 8708 Männedorf, Switzerland

## 1. Introduction

Understanding the existing social structures of a city is essential for urban planning. For many urban planning processes GIS offers key techniques for presenting and communicating the relevant socio-economic spatio-temporal data to the involved stakeholders – citizens, planners, and decision makers. Some planning aspects require sensitive data: Census, taxation, health, or ethnic data on a household or even individual level (Joung et al. 2009). This paper presents a novel approach for facing the *disclosure* dilemma in fine-grained urban planning (Armstrong et al 1999): On the one hand, it is imperative to aggregate fine-grained sensitive records large enough to avoid disclosure. On the other hand, aggregation blurs the information, potentially hindering insights about spatially explicit relationships (Leitner and Curtis, 2006). Aggregation of smaller building blocks into larger output zones appears in the literature in various contexts and is referred to as "zoning", "regionalization", "segmentation", or "partitioning" (Assunção et al. 2006, Cockings et al. 2013, Shortt 2009). Whereas recent international interest in alternatives to traditional census has produced work on generic zoning systems, zoning designs inevitably must consider specific local conceptual and practical peculiarities (Cockings et al. 2013). The here presented algorithm MASC (Multi-criteria aggregation for sensitive census data) builds a corner stone of the modularCity urban planning software environment that resulted from a Government funded innovation and development project. Its multi-criteria aggregation design specifically integrates confidentiality regulations with planning regulations and geometric requirements for the resulting zones.

## 2. Preliminaries

The input data consists of a list of parcels (building blocks) represented as two dimensional polygons and the minimum number of residents $n$ required for an area to safeguard confidentiality. The parcel attributes include the number of residents living in the buildings within its boundaries, construction regulation values (generally the maximum ratio of either floor area or building volume and plot area) and two lists of neighbors: direct neighbors are all other parcels that share a boundary, while indirect neighbors are separated by a small gap (e.g. a street). From hereafter the term reference area will be used which refers to a single parcel or a group of parcels. Initially there is one reference area for every parcel.

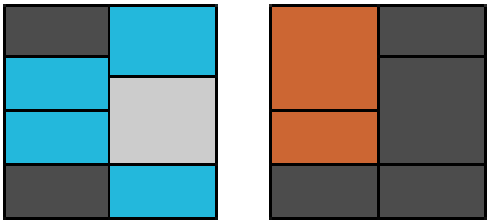


Figure 1. Neighborhood of a selected parcel (in light grey) with direct neighbors shown in blue and indirect neighbors shown in orange.

The reference areas are iteratively expanded by grouping them with a best-match neighbor based on these parameters:

- **Resident density** is the ratio of residents per area. The difference of the resident densities of both reference areas is divided by the sum of the densities for normalization, and then squared. This leads to the penalty for large relative differences, i.e. areas prefer akin areas for growing.
- **Building regulation** is the difference of the allowed floor area or building volume ratio from the zoning resolution. Here, the absolute difference is used as a penalty. This helps maintaining zones according to the building regulations, complementing pure social criteria.
- **Distance of centers** is the Euclidean distance between the centers of both polygons. This parameter favors round shapes of the growing reference areas.
- **Distance between polygons** is the length of the shortest straight line between the polygons of the two reference areas. As such it is only relevant for indirect neighbors and hinders the growing of reference areas across streets and similar gaps.
- **Anonymity penalty** is a parameter inhibiting further merging of reference areas or parcels that have already reached anonymity.

The joint development of the algorithms with local decision makers pointed out the importance of weights for the above parameters, allowing for a scenario-based planning process based on variable planning priorities. For the actual setting of the weights, any common ranking procedure can be used (Malczewski 1999). Consequently, when comparing two candidate reference areas $A_m$ and $A_n$ a difference measure is calculated as:

$$\sum \omega_i * difference_i(A_m, A_n) \qquad (1)$$

where $difference_i$ refers to the five parameters $i$ comparing the reference areas $A_m$ and $A_n$, and $\omega_i$ is the respective weight of the parameter $i$. The smaller the summed-up differences, the more similar the two reference areas are, and the more likely they will be merged.

The merging of two reference areas results in the creation of a single reference area inheriting the integrated properties of its parents, including an updated neighborhood table. The new values include the sum of the residents and building regulation figures calculated from the values of the parent areas as follows:

$$p' = \frac{p_m * a_m + p_n * a_n}{a_m + a_n} \qquad (2)$$

where $p_m$ and $p_n$ are the parameters of the to be merged reference areas $A_m$ and $A_n$ and $a_m$ and $a_n$ are their respective areas.

## 3. The MASC algorithm

The MASC algorithms starts from a list containing all reference areas that are not yet anonymous, i.e., parcels or reference areas with less than $n$ residents. The first element is removed from the list and compared to all its neighbors. For each neighbor a difference value is calculated based on the five above parameters. The neighbor with the lowest difference value, or the highest similarity respectively, is selected as the merging partner. The two reference areas are merged and the resulting reference area is, if not yet anonymous, added to the bottom of the list. The now-first element of the list is removed and the procedure is repeated until the list is empty.

```
foreach reference area p in list do
    difference = INFINITY;
    partner = null;
    foreach neighbor q of p do
        if difference(p, q) ¡ difference then
            difference = difference(p, q);
            partner = q;
        end
    end
    list.remove(p);
    list.remove(partner);
    p.mergeWith(partner);
    if !p.isAnonymous() then
        list.add(p);
    end
end
```

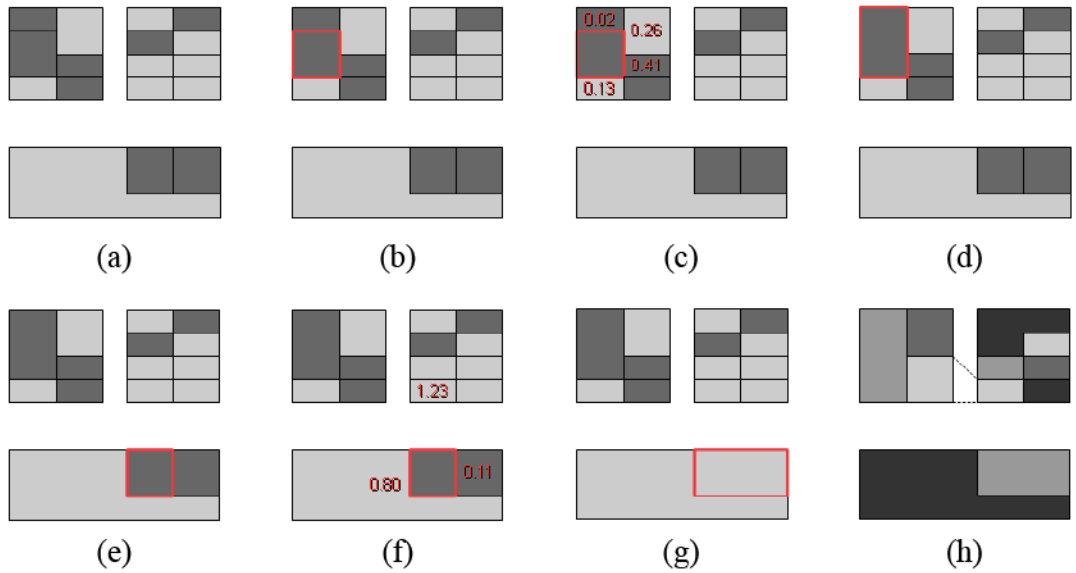



Figure 2. Process of aggregation: (a) all reference areas that are not anonymous (in dark grey) are identified; (b) a first reference area is selected; (c) the difference value with each neighbor is calculated; (d) the reference area is merged with the best match. The resulting reference area is not yet anonymous; (e) the next reference area is selected; (f) the difference value of each neighbor is calculated; (g) after merging with the best match the reference area is anonymous; (h) the 10 anonymous regions resulting from the algorithm. The dotted line shows represents the link of an reference area spanning a gap.

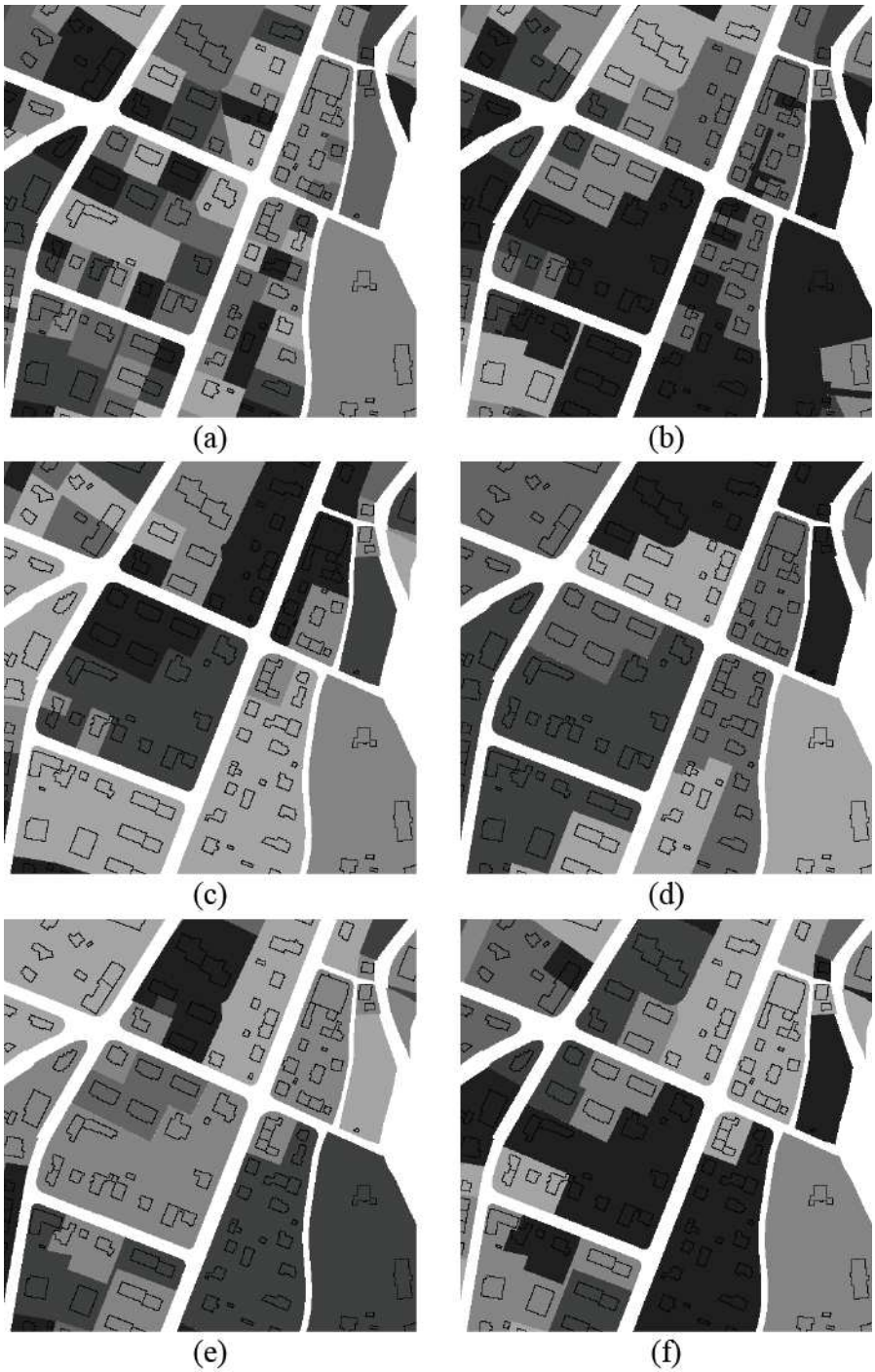



Figure 3. Different aggregation scenarios, shown for the case study of Langenthal, Switzerland. *n* was set to 30 residents (a) initial parcels that are used as the input; (b) result based on resident density only; (c) result based on floor area ratio only; (d) results based on geometric distances only; (e) result based on the following exemplary weights: density 1, floor area ratio 0.5, center distances 2.3, polygon distances 2.3, anonymity penalty 2; (f) result based on the following exemplary weights: density 1, floor area ratio 1, center distances 2, polygon distances 0, anonymity penalty 1.

## 4. Case study

The algorithm is illustrated and validated using two case studies conducted with the city of Langenthal and the smaller municipality of Männedorf in Switzerland. The most effective parameter for crystallizing the important socio-economic structure has shown to be resident density (Figure 3b). Even without any additional information about the building types (detached house, terrace, apartment building, etc.), a high weight of resident density results in zones that tend to be made up of the same residential type. When accentuating the building regulation parameter (Figure 3c), the zones are aggregated within respective building zones. Figure 3d illustrates how the distance of centers parameter inhibits the formation of inappropriately elongated slivers that conflict the notion of perceived and lived urban neighborhoods. Finally, Figures 3e and f compares two weighting scenarios illustrating the influence of the polygon distance parameter. A weight of 2.3 in Figure 3e inhibits zone growing across streets, Figure 3f with a respective weight of 0 shows several zones that are linked across streets.

## 5. Validation and Conclusions

The MASC aggregation procedure was evaluated through qualitative plausibility tests with local decision makers. In workshops with the local authorities, the stakeholders explicitly noticed that the procedure manages to reveal the fine-grained socio-economic structure whilst safeguarding confidentiality regulations. Manual adjustments were only suggested for a very small number of parcels (around 1% of all parcels). Given its weighted multi-criteria character, the procedure empowers the stakeholders to experiment with different planning scenarios adhering to different priorities (different weights). As a specific finding, the case studies revealed that – contrasting to the expectation – high-density building not necessarily led to the relocation of higher income residents. In one case study, the scenarios developed with MASC even contributed to a recent revision of the implemented zoning regulations.

## Acknowledgements

This research was supported by Swiss Commission for Technology and Innovation CTI, project 12910.1 PFES-ES, “modularCity – Software zur Unterstützung nachhaltiger Stadtplanung” and the municipality of Langenthal.

## References

Armstrong M P, Rushtong G and Zimmerman D L, 1999, Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18(5):497–525.

Assunção R M, Neves M C, Câmara G and Da Costa Freitas C, 2006, Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees, *International Journal of Geographical Information Science*, 20(7):797–811.

Cockings S, Harfoot A, Martin D and Hornby D, 2013, Getting the foundations right: spatial building blocks for official population statistics. *Environment and Planning A,* 45(6):1403–1420.

Joung C, Martins D and Skinner C, 2009, Geographically intelligent disclosure control for flexible aggregation of census data. *International Journal of Geographical Information Science*, 23(3-4):457–482.

Leitner M and Curtis A, 2006, A first step towards a framework for presenting the location of confidential point data on maps – results of an empirical perceptual study. *International Journal of Geographical Information Science*, 20(7):813–822.

Malczewski, J., 1999, GIS and multicriteria decision analysis, John Wiley & Sons, New York, NY.

Shortt N, 2009, Regionalization/zoning systems. In: Kitchin R and Thrift N (eds), International Encyclopedia of Human Geography, Elsevier, Oxford, 298–301.

# Statistical Detection of Multiple Clusters of Point Events in Small-Area Analysis Based on False Discovery Rate

R. Inoue[1]

[1]Graduate School of Information Sciences, Tohoku University, Aramaki Aoba 6-6-06, Aoba, Sendai, Miyagi 980-8579, Japan
E-mail: rinoue@plan.civil.tohoku.ac.jp

## 1. Introduction

Today, we have access to detailed location information because of improved access to statistical information, open governmental policies, and the popularisation of geo-spatial technologies. Distance-based analyses, such as the local K function (Getis and Franklin 1987), can detect clusters in point events using such information. However, these analyses cannot consider the population that affects the point event distribution. This study examines how to detect clusters using point count data aggregated on small spatial units, such as census tracts.

The major cluster detection method is the spatial scan statistic (Kulldorff 1997). This method evaluates the degree of point accumulation in the given area using a likelihood ratio. Here, this is the ratio of the likelihood of the alternative hypothesis, which assumes the given area is a cluster, to the likelihood of the null hypothesis, which assumes the area is not a cluster. The area with the maximum likelihood ratio is searched, and then its significance is tested by comparing it to the distribution of the maximum likelihood ratio obtained from the Monte Carlo simulation. The spatial scan statistic and its derivations are widely used. However, they are limited when detecting multiple clusters, because the alternative hypothesis presumes the existence of a single cluster, and multiple testing should be avoided. Although subsequent clusters can be detected under the condition that previously detected clusters exist at the detected locations, this limitation spoils the availability of spatial scan statistics-based method.

Recently, two approaches for multiple cluster detection have been proposed. The first is an extension of the spatial scan statistic (Mori and Smith 2010), and the second applies the false discovery rate (FDR) controlling procedure (Brunsdon and Charlton 2011). However, applying these methods to a small-area analysis is problematic. The first method takes a long time to calculate, even when there are few regions, and the second method may lose statistical power when the regions become small. Therefore, this study proposes a new method to statistically detect multiple clusters in a small-area analysis by combining these two approaches, and then checks its applicability.

## 2. Previous Approaches to Multiple Cluster Detection

Mori and Smith (2010) proposed evaluating multiple cluster models using the Bayesian information criterion (BIC) as an expansion of the spatial scan statistic. This method first forms ‘cluster schemes’ that set the locations of multiple cluster candidates. Then it estimates the density parameters based on the assumption of point distribution and calculates the BICs. After selecting the cluster scheme with the maximum BIC, the significance of the scheme is tested using a Monte Carlo simulation. This is a promising method for multiple cluster detection, since the model selection using the BIC can consider the number of clusters and their locations. However, since the number of possible cluster schemes is huge and no efficient search procedure is proposed, it might take an excessive amount of time to detect clusters. For example, the

process may take as long as a week for around 2,000 spatial units. This limitation is caused by the global search procedure of cluster schemes, in which the locations of all cluster candidates are needed when calculating the BIC.

Brunsdon and Charlton (2011) proposed using the FDR controlling procedure, which can avoid the multiple testing problem and achieve greater statistical power than the familywise error rate controlling methods (e.g. Holm 1979). The FDR controlling method works as follows. Suppose m hypotheses are tested and R null hypotheses are to be rejected (see Table 1). Multiple testing increases the occurrence of type I errors (V) by chance. Benjamini and Hochberg (1995) defined the FDR as an index of false discoveries,

$$\mathrm{FDR} = \mathrm{E}\left(\frac{\mathrm{V}}{\mathrm{R}}\right), \quad (\mathrm{FDR}=0 \text{ if } \mathrm{R}=0) \tag{1}$$

and proposed an FDR controlling procedure that keeps the FDR less than the given significance level, $\alpha$. Brunsdon and Charlton (2011) use this for cluster detection. The method configures the set of alternative hypotheses that each spatial unit is a cluster. Then it tests them using the FDR controlling procedure. It is a simple statistical approach, but it might lose statistical power when the spatial units are small. Therefore, it is necessary to form larger cluster candidates when analysing small-area data by combining spatial units.

Table 1. The m hypotheses test.

| | Rejected null hypothesis | Retained null hypothesis | Total |
|---|---|---|---|
| Null hypothesis is true | V | U | $m_0$ |
| Alternative hypothesis is true | S | T | $m - m_0$ |
| Total | R | m – R | m |

## 3. A New Multiple Cluster Detection Method

The proposed method in this study conducts a local search of cluster candidates, which are combinations of adjacent spatial units. The method then evaluates candidates using the FDR controlling method.

The proposed procedure is as follows. First, all spatial units are added to the list of candidates. Then, combinations of adjacent spatial units are created. Including units with smaller than average densities always increases the p-value of the null hypotheses. Therefore, the method only considers spatial units with higher than average densities. A spatial unit is randomly selected as a seed for the local search. Then, the combination of units adjacent to the seed unit that outputs the lowest p-value of the null hypotheses is added to the list of the candidates. This combination becomes the seed for the subsequent candidate formation. This process is repeated until no combination with a lower p-value can be formed.

Once the local search of cluster candidates finishes, the significance of each potential cluster is tested using the FDR controlling procedure. Since the created cluster candidates are not independent, their significance is tested based on the work of Benjamini and Yekutieli (2001), who proposed a method to test dependent hypotheses.

## 4. Application

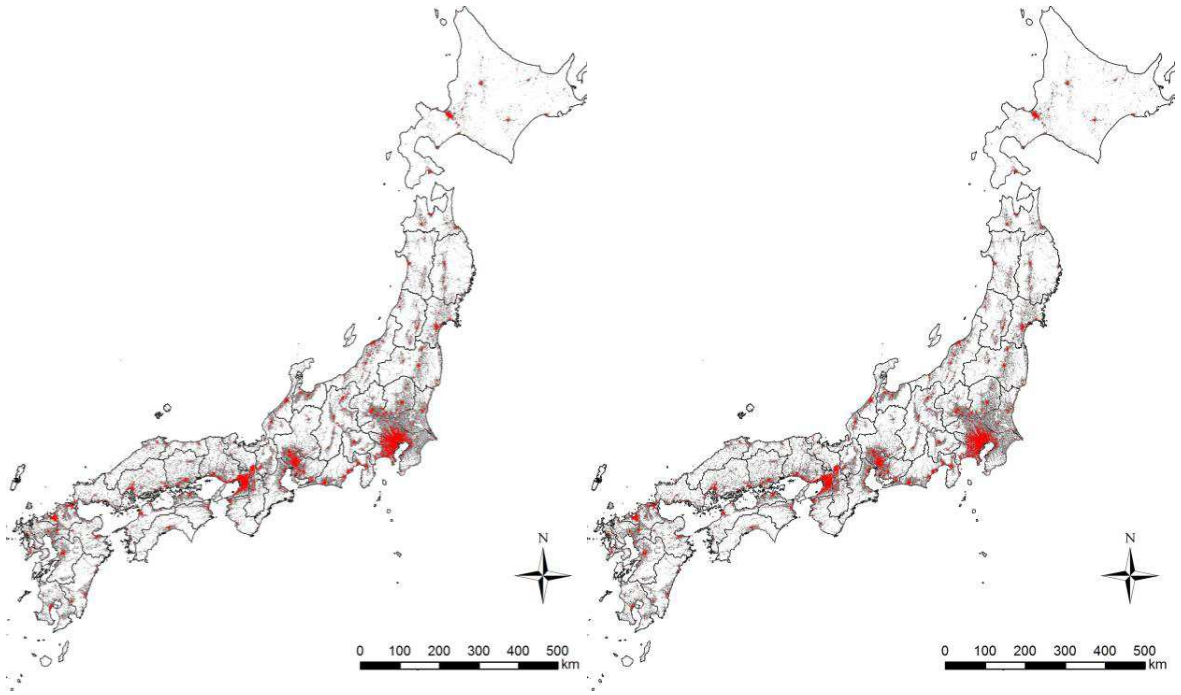
The proposed method is compared to the method of Brunsdon and Charlton (2011) by applying the methods to two datasets with different size spatial units: 500 m grids and links in a road network. Both applications use an FDR significance level of less than 1%.

The first application counts the offices of all industries on 500 m grids obtained from the '2009 Economic Census for Business Frame' of Japan. In all, 6,043,300 offices are located on 336,451 grids. The red areas in Figure 1 illustrate the clusters detected by the proposed method and the Brunsdon–Charlton method. The grey areas illustrate non-cluster grids that contain offices. The proposed method selects 45,016 grids, while the Brunsdon–Charlton method selects 44,696 grids. There is little difference because the spatial units are not small enough. The proposed method took two minutes to obtain its results on a computer with a Xeon 2.5 GHz CPU.

The second application counts restaurants on road network links in the central three wards of Tokyo. The locations of 2,972 restaurants are taken from the 'Telepoint Pack Database,' the telephone directory of February 2011. The 1,115 km of road network, comprising 19,338 links, is taken from the 'Digital Map 25000 (Spatial Data Framework)', published by the Geospatial Information Authority of Japan. Figure 2 illustrates the detected clusters. The proposed method detects 1,173 links, while the Brunsdon–Charlton method detects only 274 links. This confirms that the proposed method has an advantage in terms of statistical power when the spatial units are small. The proposed method executed in less than a second on a computer with a Xeon 2.5 GHz CPU.

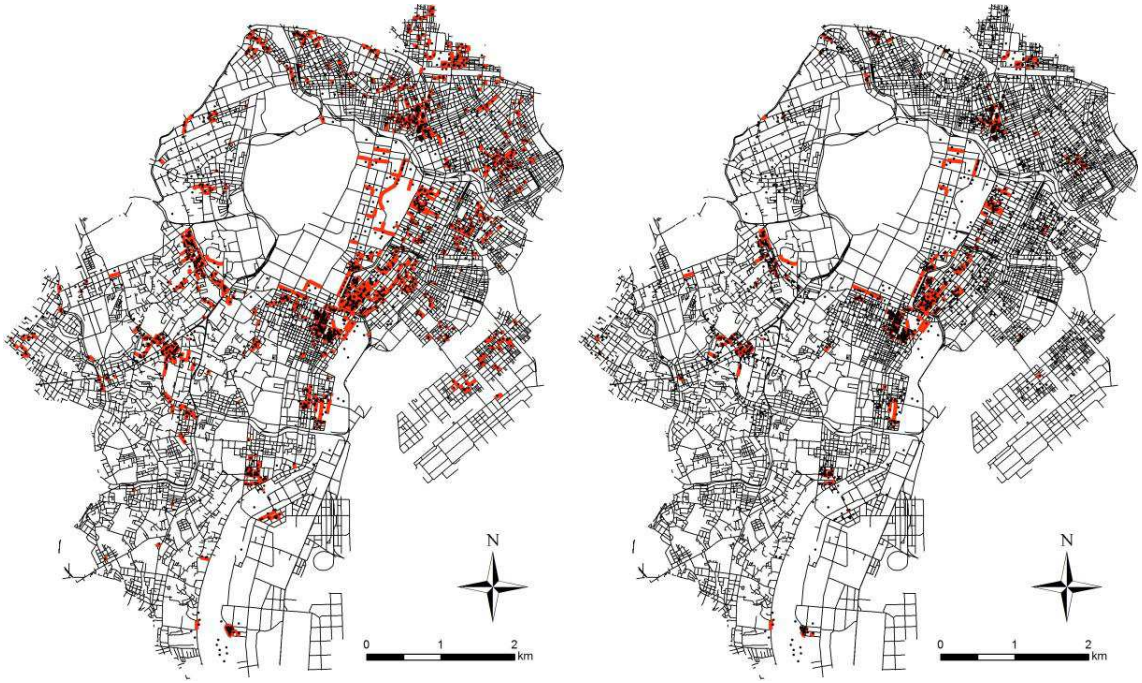
## 5. Conclusion

This study proposes a method for multiple cluster detection based on the FDR. Tests reveal that the proposed method succeeds in detecting clusters of small regions in a short time.




(a) Proposed method (b) Brunsdon–Charlton method

Figure 1: Detected clusters of offices (a part of Japan).




(a) Proposed method (b) Brunsdon–Charlton method

Figure 2: Detected clusters of restaurants.

## Acknowledgements

This work was supported by JSPS KAKENHI, grant number 24241053. The 'Telepoint Pack Database' provided by Zenrin Co., Ltd. and the '2009 Economic Census for Business Frame' provided by Sinfonica were used as part of the CSIS Joint Research (No. 456) of the Centre for Spatial Information Science, University of Tokyo.

## References

Benjamini Y and Hochberg Y, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57(1):289–300.

Benjamini Y and Yekutieli D, 2001, The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29:491–507.

Brunsdon C and Charlton M, 2011, An assessment of the effectiveness of multiple hypothesis testing for geographical anomaly detection. Environment and Planning B: Planning and Design, 38:216–230.

Getis A and Franklin J, 1987, Second-order neighborhood analysis of mapped point patterns. Ecology, 68(3):473–477.

Holm S, 1979, A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2):65–70.

Kulldorff M, 1997, A spatial scan statistic. Communication Statistic Theory and Method, 26(6):1481–1496.

Mori T and Smith T, 2010, A probabilistic modeling approach to the detection of industrial agglomeration. KIER Discussion Paper, 777:1–54.

# Location Prediction With Sparse GPS Data

Ayush Jaiswal[1], Yao-Yi Chiang[2], Craig A. Knoblock[3], Liang Lan[4]

[1]National Institute of Technology Calicut, India
Email: ayush_bcs10@nitc.ac.in

[2]Spatial Sciences Institute, University of Southern California, USA
Email: yaoyic@usc.edu

[3]Department of Computer Science and Information Sciences Institute, University of Southern California, USA
Email: knoblock@isi.edu

[4]Noah's Ark Lab, Huawei Technologies
Email: lan.liang@huawei.com

## 1. Introduction

Predicting the next location of a user from their movement history is useful in building intelligent applications that can continuously assist users without explicit user-input. Data collected by applications on consumer-grade mobile devices, such as GPS data, can have missing records (e.g., due to the application crashing) and the sensor sampling frequency needs to be kept low so that it does not drain the mobile battery. Thus, there can be a significant time gap between each pair of recordings. Our work in this paper focuses on predicting the next location of a mobile user using such sparse GPS data, collected at a very low frequency of once every 10 minutes. To give an example of dense data, Krumm and Horvitz (2005, 2006) use data collected once in every six seconds.


Figure 1: Movement patterns may be disjoint. The blue and the red points were recorded on two different days.

Sparseness in GPS data makes finding patterns in a user's movement history difficult. Moreover, the low sampling rate might capture movement patterns that are along the same path but are disjoint (Figure 1). Losses in GPS connection and imperfect behavior of the data collection application further increase the sparseness of the data. We tackle the problem of sparseness by interpolating user movements using a routing service.

Location prediction can be viewed as a classification problem in which the possible next locations are discrete classes, but GPS data is continuous in nature. Hence, we use a grid over the region where the data is centered, and map the points to grid-blocks. Another possible method of location abstraction is mapping points to the nearest mapped addresses according to maps such as Baidu Maps and OpenStreetMap. This is known as reverse geocoding. This

approach depends significantly on the accuracy and the amount of address information available for the region where the data is collected. With insufficient address information, such as in our case, using reverse geocoding results in a lot of repetition in location-IDs as many points map to a single location-ID. This leads to loss of movement information.

We also discuss the results of using four different Markov models for the prediction task on the sparse and the processed data.

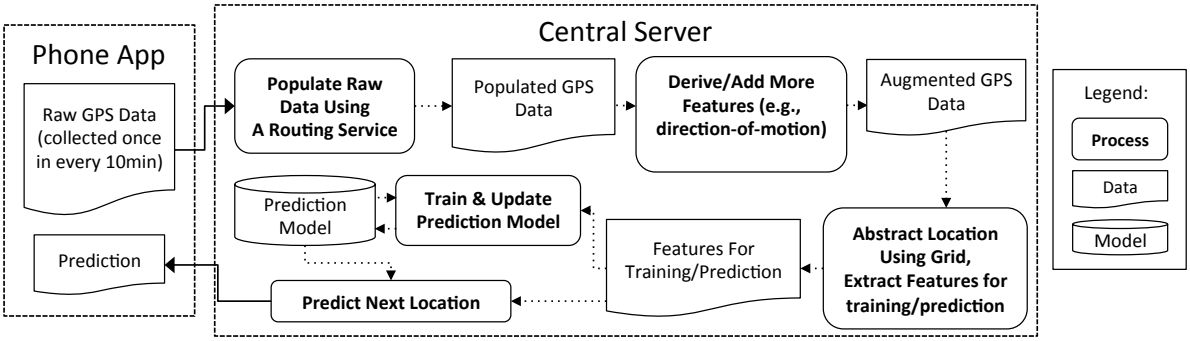



Figure 2: Location Prediction System for Sparse GPS Data.

## 2. Next Location Prediction

Figure 2 shows the overall workflow of our approach. The sparse GPS data is populated using a routing service to produce a dense set of user movement history, additional features (such as direction-of-motion, described later) are added, and the points are abstracted to locations using a grid. The resulting features are given as inputs to the prediction model.

### 2.1 Dealing with Sparseness

Our approach uses a routing service to find the shortest path between every consecutive pair of points and uses the route returned to fill the gap between the pair with dummy points. The underlying assumption is that people tend to take the shortest path between any two places that are near one another, especially when they are separated by just 10 minutes in time.

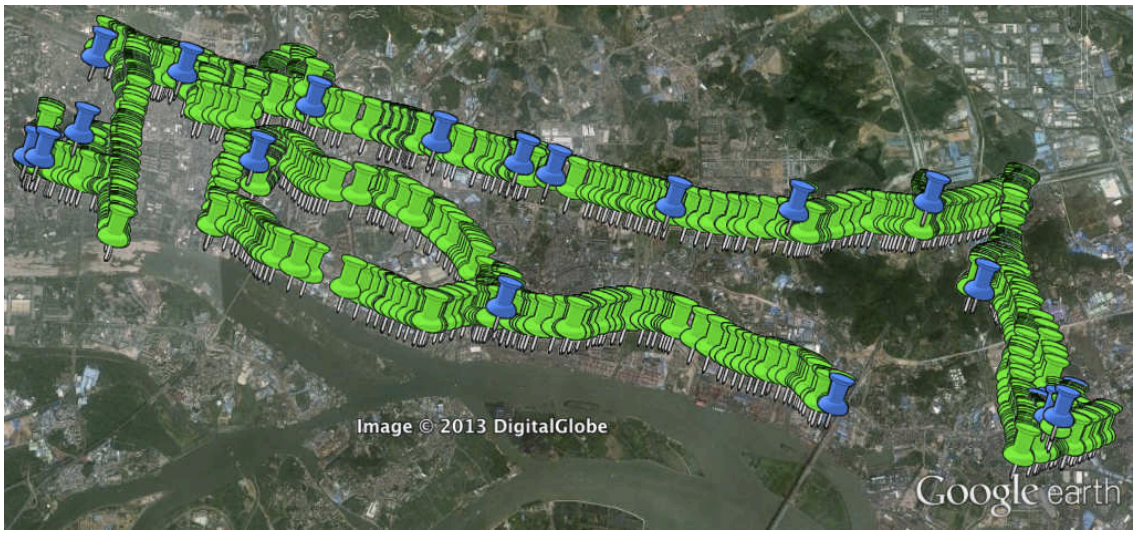



Figure 3: The blue points are original points in the data while the green ones were added using the routing service.

For example, Figure 3 shows how our system populated some of the data that we work on. The interpolated points filled in using the routing service complete the original path very

elegantly. We use the Google Directions API[1] to get the shortest driving route between consecutive pairs of points.

### 2.2 Features and Prediction Models

We use Markov models to predict the next grid-block the user will be in, as illustrated in Figure 4. Markov models help in describing sequences of events and their associated probabilities. Cheng et al. (2003) explain how Markov models can be used for location prediction. We employ four different Markov models to test four hypotheses for location prediction from sparse GPS data:

- order-1 Markov model (O1MM): predict the next location of the user based on *their last known location*
- order-2 Markov model (O2MM): predict the next location based on *their two last known locations*
- order-2 Markov model with fallback on order-1 Markov model (FMM): try predicting with O2MM, and when it fails to make a prediction, use O1MM
- order-1 Markov model with direction-of-motion feature (O1MMD): we use the direction-of-motion between every consecutive pair of points. The directions that we employ are: North, North-East, East, South-East, South, South-West, West, North-West, and *stationary*. This feature removes the need of keeping track of multiple previous locations as it captures the information contained in them.

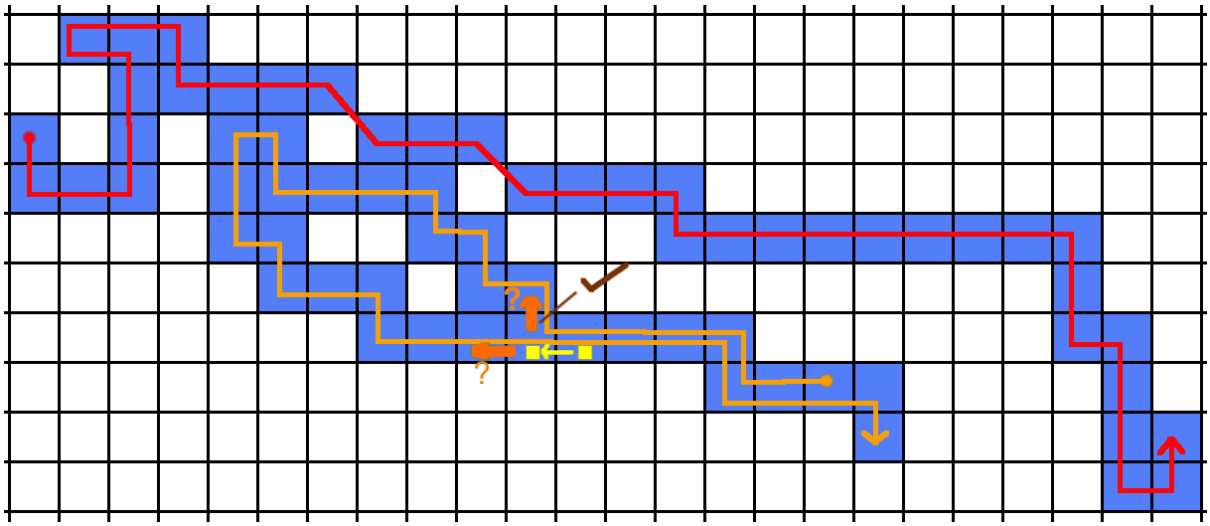


Figure 4: Predicting the next grid-block the user will be in. The model has learnt user movement patterns from day-1 (red line) and day-2 (orange line). On day-3 (yellow squares), it predicts the next location of the user in the upward direction (as learnt from the previous day).

## 3. Experiments and Results

Our data were collected by a user in Shenzhen, China over a 24 day period. On average, it has 14 GPS points in a day. We used the aforementioned Markov models for the task of location prediction on both the original data and the data resulting from the application of our processing steps. We calculated the average prediction accuracies using two experiment settings: the leave-one-day-out cross-validation setting (L1CV) uses the data from a particular day as test data and data from all other days as training data, and the sequential data

[1] http://developers.google.com/maps/documentation/directions

setting (SEQ) that uses data from a particular day as test data and data from only the days in the movement history before that day as training data. While cross-validation is a general approach to comparing the accuracies of machine learning models, SEQ is closer to how we would want the prediction to work in real world settings. A correct prediction is one that matches the next observed grid-block of the user. Our accuracy measure is the fraction of predictions that are correct.

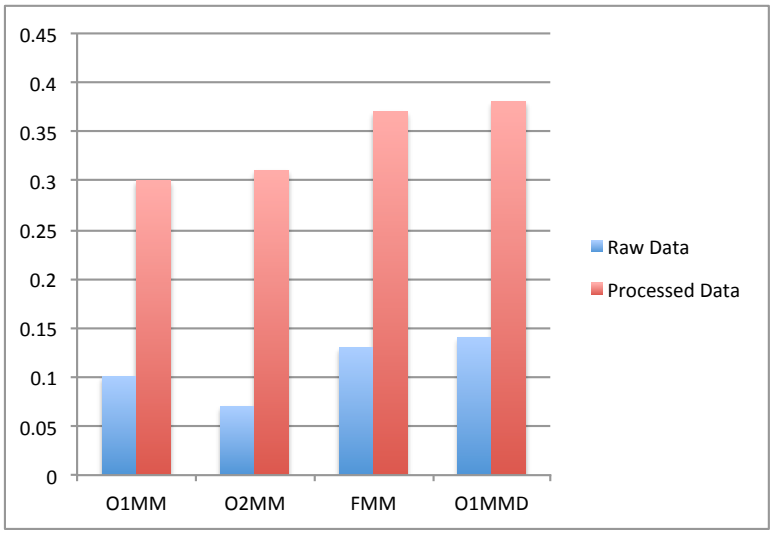



Figure 5: Average SEQ accuracy.

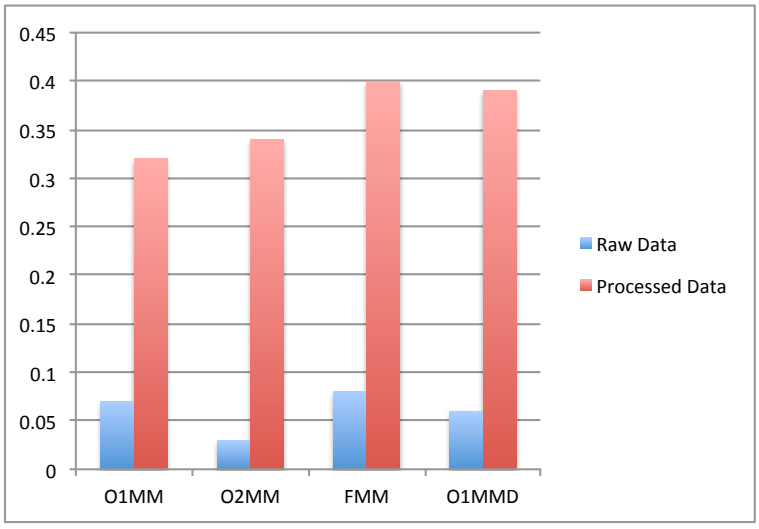



Figure 6: Average L1CV accuracy.

Figures 5 & 6 summarize our results. O1MMD and FMM performed almost equally well and better than the other models on the processed data. The desired order of accuracies should be O1MM ≤ O2MM ≤ FMM as the ones to the right make use of more information about the user's history, but we do not find this order in case of sparse data as O2MM could not learn many patterns because of the sparseness. In general, the prediction models were unable to learn patterns in the user's movements from the sparse GPS data. Solving the problem of sparseness improves their prediction accuracies. The overall accuracies appear low because of significant randomness in the movement patterns of the user whose data we used. It has been found that randomness in a user's movement patterns reduces the accuracy of prediction models (Anagnostopoulos et al. 2009). Such randomness is inevitable in the movements of real users.

## 4. Related Work

Krumm and Horvitz (2006) use grid-based location abstraction to predict the destination of the user from partial trajectories. Our work is different from theirs as we predict the user's next location, and our data is much more sparse than theirs. While their data is collected once every 6 seconds, ours is collected once every 10 minutes. Gao et al. (2012) report that Hierarchical Pitman-Yor language gives a higher accuracy as compared to Markov models. Anagnostopoulos et al. (2009) implemented location prediction using decision trees, k-nearest neighbor, and ensemble learning algorithms, and found that ensemble learning algorithms performed the best among them. The methods proposed in this past work cannot be applied directly to sparse data, such as ours, as the machine learning algorithms used in them will be unable to learn patterns effectively. Our processing steps interpolate the sparse data and improve location prediction on such data.

## 5. Discussion and Future Work

This paper presented an approach for location prediction using sparse user movement history. We showed that by exploiting an online routing service, we made location prediction possible on sparse movement data. We plan to build an intelligent method for automatically generating the dynamic grid size specific to a dataset and to incorporate other sensor data on mobile phones into the location prediction framework.

## Acknowledgements

We thank the USC Viterbi School of Engineering for providing summer fellowship to Ayush Jaiswal, and we thank Huawei Technologies Co. Ltd. for a gift that supported the project and for providing the data.

## References

Christine Cheng, Ravi Jain and Eric van den Berg, 2003, Location prediction algorithms for mobile wireless systems. *Wireless Internet Handbook, 245-263, CRC Press, Inc. Boca Raton, FL, USA*.

Huiji Gao, Jiliang Tang and Huan Liu, 2012, Mobile location prediction in spatio-temporal context. *Proceedings of the Mobile Data Challenge by Nokia Workshop in conjunction with the International Conference on Pervasive Computing, Newcastle, U. K*.

John Krumm and Eric Horvitz, 2005, The Microsoft Multiperson Location Survey. *Microsoft Research (MSR-TR-2005-103): Redmond, WA USA*.

John Krumm and Eric Horvitz, 2006, Predestination: Inferring Destinations from Partial Trajectories. Proceedings of the *8th International Conference on Pervasive and Ubiquitous Computing (UbiComp) 243-260*.

Theodoros Anagnostopoulos, Christos Anagnostopoulos, Stathes Hadjiefthymiades, Miltos Kyriakakos and Alexandros Kalousis, 2009, Predicting the Location of Mobile Users: A Machine Learning Approach. *ICPS '09 Proceedings of the 2009 International Conference on Pervasive Services, 65-72*.

# Integrated geons: spatial-explicit modelling of latent phenomena

S. Kienberger[1], M. Hagenlocher[1], S. Lang[1]

[1]Interfaculty Department of Geoinformatics – Z_GIS, University of Salzburg, Schillerstrasse 30, 5020 Salzburg, Austria
Email: {stefan.kienberger; michael.hagenlocher; stefan.lang}@sbg.ac.at

## 1. Introduction

Much endeavour is currently given to represent as well analyse and understand complex, latent phenomena. Latency (lat. for abeyant or unseen) is characterised through a certain existence which however did not materialise itself into directly measurable outcomes. This may sound philosophical, but is a key aspect to be considered when modelling multi-dimensional phenomena in space (and time). An example is risk to a given threat, which is characterised by single or multiple hazards and the latent vulnerability of society or differential population groups. Risk manifests itself into direct, measurable impacts when a hazardous event occurs. However, latency is not only a characteristic of risk, but also implicit to other complex phenomena, such as quality of life, human well-being or landscape sensitivity. These latent phenomena share a clear policy-relevance. Their assessment and monitoring over space and time should guide decision makers to define, implement and evaluate context and place-specific interventions.

This paper discusses a novel approach to represent and operationalize such phenomena in a spatial-explicit manner. We thereby refer to the concept of geons that was introduced by Lang et al. in 2008, and which has been expanded to the concept of integrated geons (Lang et al., 2014). We elaborate on how integrated geons are conceptualised and modelled. While key methods to model integrated geons have been established for a while (Kienberger et al, 2009), we discuss opportunities for an entirely spatially enabled workflow. To evaluate the benefits and challenges, we compare the outputs of an integrated geon approach with results applying traditional approaches. Finally, we discuss future pathways on how integrated geons can be categorized based on certain typologies or qualities. The paper presents ongoing work, and is intended to stimulate discussions.

## 2. The systems perspective – (integrated) geons

Recently, Lang et al. (2014) defined geons as spatial objects, which are homogenous in terms of varying spatial phenomena under the influence of policy intervention and are generated by scale-specific spatial regionalization of a complex, multidimensional geographical reality incorporating expert knowledge. In this paper we follow the concept of integrated geons, which addresses abstract, yet policy-relevant phenomena.The geon concept abstracts from the level of underlying data sets towards a domain-specific representation by adapting the scale of the constructed geons to that of policy intervention. The goal is to generate key reference units for policy-related action. Geons show uniform response regarding the spatial phenomenon under concern, and are expert validated regarding usability and relevance (Lang et al. 2010). They integrate spatial information and exhibit emergent properties of systemic areal units. In congruence with systems thinking (Laszlo 1972), we argue that geons bear a dual character sensu Koestler (1967) in terms of hierarchical organization. The nested behavior can be applied to systemic areal units, as a geographical correspondence to holons (Wu 1999). From that perspective, geons are considered not only integrated wholes, but

likewise as parts of a larger spatial (latent) phenomenon. To achieve a spatial representation of the latter, we try to map these parts in a spatially explicit way. Geons integrate a set of geospatial information layers in a way to 'give message' (Allen & Starr 1982) on a policy level, as an exhaustive representation of latent phenomena.

### 2.1 Modelling integrated geons – pitfalls of current approaches

The workflow comprises five major modelling stages. These include (1) the definition of a hierarchical conceptual framework for systematizing the latent phenomenon of interest; (2) indicator selection and data pre-processing; (3) weighting and aggregation through regionalization; (4) an assessment of the sensitivity/robustness; and finally (5) visualizing the results (Lang et al., 2014). Different statistical pre-processing routines are used to statistically define a sound set of indicators. This includes for instance the reduction of existing multicollinearities in the data. Regionalization techniques segment the continuous datasets in an n-dimensional indicator space, where n is the number of input datasets. A workflow for modelling integrated geons was initially proposed by Kienberger et al. (2009) and has since then been refined and transferred to other application domains.

Although suitable per se, several of the traditional pre-processing methods are not optimized for continuous geospatial datasets. Winsorization, for example, is a standard procedure to treat extreme values. However, this approach does not take into account the spatial characteristics and topological relationships of geospatial datasets. The same holds true for weights assigned to individual datasets. Most approaches assume an equal relative importance of a single indicator for the entire study area, which might, not be the most appropriate approach, as relationships between the phenomenon to be modeled and its underlying indicators may vary in space. We therefore propose using spatial explicit approaches for data pre-processing and spatial differential weighting of the data, such as low-pass filters for outlier treatment, or geographically weighted regression (GWR) for spatial explicit weighting.

## 3. The benefits and challenges of integrated geons – a comparison with composite indicator approaches and Multi Criteria Assessment (MCA)

The concept and method to delineate integrated geons expands beyond traditional methods. To be able to discuss benefits and challenges, we compare presently modelled social malaria vulnerability units (as an instance of integrated geons) for the same study area and with the same underlying datasets with approaches (1) using an administrative boundary based composite indicator and (2) a grid-based multi criteria assessment (MCA). Based on that, we reflect on the benefits and challenges of integrated geons. For instance we believe that administrative based composite indicators - usually visualized as choropleth maps - are not spatial in a stricter sense, as no spatial methods are applied besides its visualization. Additionally this approach is strongly influenced by the Modifiable Areal Unit Problem (MAUP). On the other hand, grid-based approaches provide a much more spatial-explicit representation, but do not reflect specifically scale dependent instances and may be limited in representing hierarchical relationships. A major advantage beyond MCA is that each geon is not only defined by its index value based on the vector magnitude but additionally through its relative and absolute contribution of the underlying indicators. This enables decision makers to identify context and place-specific interventions in a spatial explicit way. Furthermore, this provides the opportunity to categorize the resulting geons based on its similarity in contributing indicators. Currently methods are being explored to achieve that. Preliminary results are based on spatially enabled cluster analysis. Challenges of integrated geons evolve

around the novelty of its approach and the need for new spatial pre-processing and global sensitivity analysis methods. Additionally, positive feedback has been expressed by a variety of different users. However, an empirical study is needed to proof its usefulness based on a set of defined evaluation criteria.

## 4. Conclusions

We defined integrated geons in a systemic and holistic way with a clear policy relevant purpose. We highlighted challenges in the workflow and propose spatial explicit methods to be included in current pre-processing workflows. Additionally, we explored the benefits and challenges through a case study comparing results to traditional approaches. In summary, the proposed approach is an opportunity to spatially represent complex latent phenomena with the final objective to match it to the respective policy scale of different users.

## Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 266327 (HEALTHY FUTURES, http://www.healthyfutures.eu/) and from the Austrian Science Fund (FWF) through the Doctoral College GIScience (DK W 1237-N23).

## References

.

Allen TFH and Starr TB, 1982, Hierarchy. University of Chicago Press., Chicago

Kienberger S, Lang S and Zeil P, 2009, Spatial vulnerability units – expert-based spatial modelling of socio-economic vulnerability in the Salzach catchment, Austria. Nat. Hazards Earth Syst. Sci., 9:767-778

Koestler A, 1967, The ghost in the machine. Hutchinson, London.

Lang S, Zeil P, Kienberger S and Tiede D, 2008, Geons – policy-relevant geo-objects for monitoring high-level indicators. In: Car A, Griesebner G and Strobl J (eds.), *Geospatial Crossroads @ GI_Forum '08.* Proceedings of the Geoinformatics Forum Salzburg, 180-185

Lang S, Albrecht F, Kienberger S and Tiede D, 2010, Object validity for operational tasks in a policy context. Journal for Spatial Science, 55 (1): 9-22

Lang S, Kienberger S, Tiede D, Hagenlocher M and Pernkopf L, 2014, Geons – domain-specific regionalization of space. Cartography and Geographic Information Science, 41(3):214-226

Laszlo E, 1972, The systems view of the world. New York, George Braziller

Wu J, 1999, Hierarchy and scaling: extrapolating information along a scaling ladder. Canadian Journal of Remote Sensing, 25:367-380

# meteo: package for automated meteorological spatio-temporal mapping

Milan Kilibarda[1], Branislav Bajat[1],Tomislav Hengl[2], Milutin Pejović[1]

[1]University of Belgrade, Faculty of Civil Engineering, Department of Geodesy and Geoinformatics, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia
Email: {kili; bajat; mpejovic}@grf.bg.ac.rs
[2]ISRIC - World Soil Information P.O. Box 353, 6700 AJ Wageningen, the Netherlands
Email: tom.hengl@wur.nl

## 1. Introduction

The most powerful *R* (R Development Core Team 2012) package available for geostatistical analysis is *gstat*, which was developed for applied geostatistics (Pebesma 2004). Many spatial geostatistics techniques (including ordinary, universal kriging, block kriging, kriging in a local neighborhood, variogram cloud diagnostics, variogram modeling, multivariable variogram modeling, cokriging and simulation) are available to the broad community of geoscientists. The development of the *spacetime* package has already started in 2010 and the *gstat* functions have been adapted for spatio-temporal mapping (Pebesma 2012), including spatio-temporal variogram fitting and implementation of global spatio-temporal ordinary kriging.

Hengl et al. (2012) describe a framework for space-time regression kriging interpolation of daily temperatures that makes use of a time-series of MODIS images, which are presented as a Croatian case study. Kilibarda et al. (2014) made spatio-temporal interpolation for the mean, maximum and minimum temperature using spatio-temporal regression-kriging with a time series of MODIS 8 day images, topographic layers (DEM and TWI) and a geometrical temperature trend as covariates. The model and predictions were built for the year 2011 only, for the global land areas, but the same methodology could be extended for the whole range of the MODIS LST images (2001–today). Global spatio-temporal variograms and regression models described by Kilibarda et al. (2014) are stored in the *meteo R* package for the purpose of automated mapping of daily temperatures at 1 km/ 1 day resolution.

This article describes the *R* package *meteo* that is still under development. The package provides functionalities for the automated mapping of meteorological observations using spatio-temporal regression kriging. The automated spatio-temporal kriging interpolation procedure is a data driven approach designed for mapping with little or no human interaction. Currently, automated mapping with the *meteo* package can be decomposed in chunks:

1. defining input observations and covariates;
2. use of pre-calculated global models;
3. detecting and/or removing outliers;
4. creation of final prediction (and its export to GIS formats);
5. cartographic visualisation of results and/or creation of web maps (e.g. by using R package *plotGoogleMaps* (Kilibarda and Bajat 2012) for automatic creation of interactive web maps).

In addition, *meteo* offers the possibility of using user defined covariates, regressions and variograms; thereby giving more flexibility of using the package in a semi-automated approach.

## 2. Implementation

The *R* is a system for statistical computation and graphics, which provides programming facilities, high-level graphics, interfaces to other languages, and debugging facilities (R Development Core Team 2012). *R* is free and open source software under the terms of the GNU General Public License. *R* is organized as a collection of packages designated for specific tasks.

The package *meteo* has been implemented in the R environment for statistical computing. It combines functionalities of the *rgdal* (*GDAL* raster/*OGR* vector data import/export), *raster* packages (raster data loading and analysis), *spacetime* (classes and methods for spatio-temporal data), *gstat* (geostatistics) and *snowfall* package (cluster computing). Spatio-temporal regression kriging prediction and cross validation have been implemented in *meteo*, and presumably it has not been implemented in any other software yet. The set of the so far created functions of *meteo* package is given in Table 1, package *meteo* is available on https://r-forge.r-project.org/projects/meteo/, under GPL licence.

Table 1. The functions in *meteo* package.

| Function | Description |
|---|---|
| *meteo2STFDF* | Creates an object of STFDF (spatio-temporal data with full space-time grid) class from two data frames (observation and stations). The observations data frame contains at least: station ID column, time column (day of observation) and measured variable column. Stations data frame contains at least: station ID column, longitude (or x) and latitude (or y) column. |
| *rm.dupl* | This function finds point pairs with equal spatial coordinates from STFDF object and remove locations with fewer observations. |
| *tgeom2STFDF* | Calculate geometrical temperature trend for mean, minimum or maximum temperature. (see Kilibarda et al. 2014). |
| *tiling* | Tiling raster or Spatila Grid/Pixels object to smaller parts with optional overlap. |
| *pred.strk* | Function for spatio-temporal regression kriging prediction based on *gstat* krigeST function (global spatio-temporal ordinary kriging). |

Function for spatio-temporal regression kriging prediction (*pred.strk*) in *meteo* package applies a tiling procedure for prediction. The area is divided into tiles (smaller parts) by the *tiling* function, which is implemented in the *meteo* package. For each tile, the nearest spatio-temporal observations are selected according to distance from tile's centroids. Subsequently, spatio-temporal regression kriging estimates values within each tile on the base of nearest selected observations. Thus, within each tile, all estimates are calculated by using global kriging from previously selected observations. In contrast to traditional kriging in the local neighborhood approach, applied algorithm reduces the number of spatial searching for nearest observations, coming up to one search per tile, instead one per each location.

## 3. Case study: Automated mapping mean daily temperatures in Serbia

The collection of stations from Global Surface Summary of Day and European Climate Assessment & Dataset data were used for mapping the mean daily temperatures in Serbia from 2011-07-05 to 2011-07-08. Observation data (for July 2011) are stored in the *meteo* package as table data (*data.frame*) for the purpose of demo examples. The corresponding spatial information are stored in the package as the same class.

The covariates for Serbia (2011-07-05 to 2011-07-08) are stored in the package including two dynamic covariates geometrical temperature trend and splined MODIS LST (see Kilibarda et al. 2014), as well as two static covariates DEM and TWI. Figure 1 shows a spatio-temporal plot of the splined MODIS LST over the domain of interpolation.

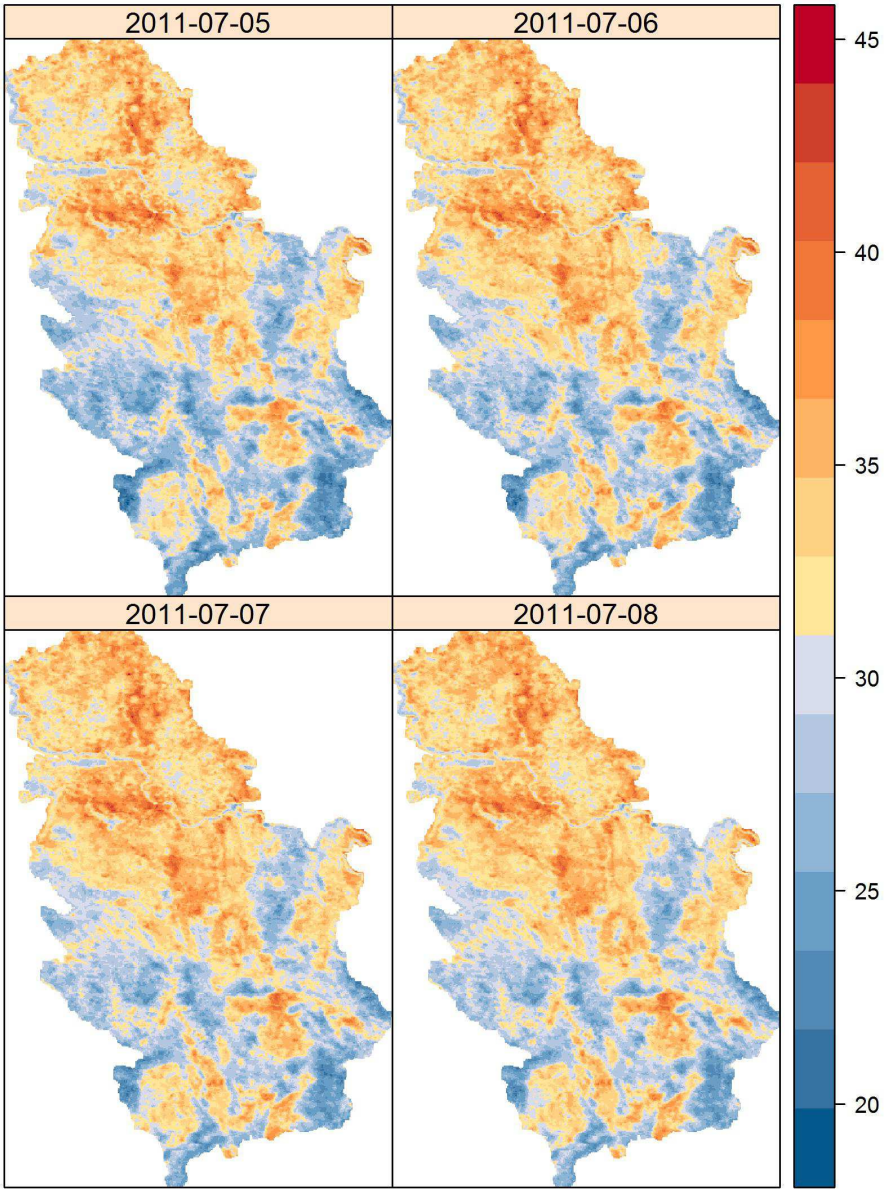



Figure 1. Splined MODIS LST 8-day images in Serbia (2011-07-05 to 2011-07-08).

The prediction of mean daily temperature (Figure 2) was produced based on the observed data of only 27 stations. The trend part was computed by regression model built in the function (methodology is described in detail in Kilibarda et al. 2014) using previously described covariates. MODIS LST images (Figure 1) are significant estimators for mean air

temperatures, despite the evident big difference between land and air temperatures, which is typical during the summer.

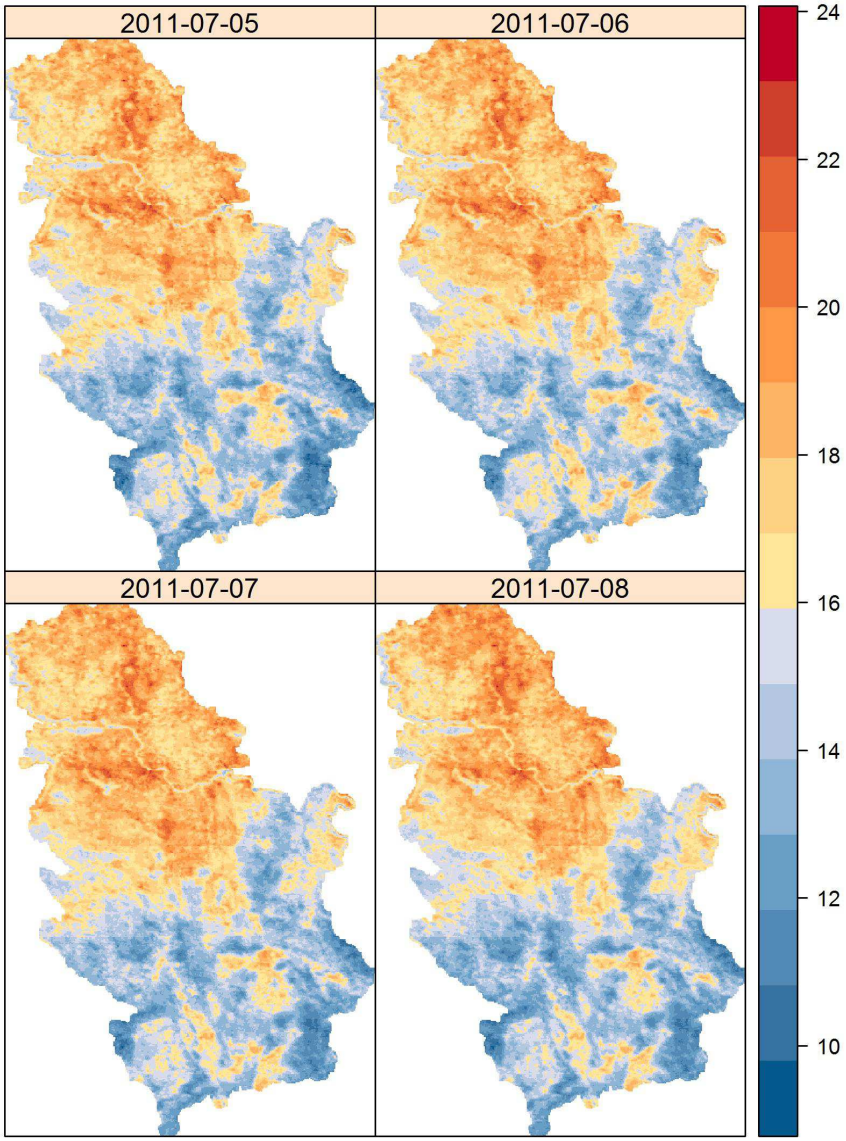



Figure 2. Prediction of mean daily temperature for Serbia (from 2011-07-05 to 2011-07-08) produced by automated mapping.

## 3. Discussion and conclusion

The illustrated mapping framework enables the use of spatio-temporal regression kriging for meteorological mapping. The implementation of the fast searching algorithm provides an advantage in computing when completing interpolations over a large spatio-temporal grid. The advantage is especially noticeable in case of the grids containing longer time series (e.g. predictions made for the area of interpolation over a year period where each location contains around 365 observations).

The automated mapping framework presented herein is still under development and a lot of functionalities need to be implemented in the future. There are still many open questions related to a) an optimal number of tiles for the domain of interpolation, b) the choice of the

optimal threshold for the automated detection of outliers, and c) incorporating the function for downloading ground station observations from data providers. Likewise, the development of procedures for downloading and mosaicking remote sensing imagery and their organization in an appropriate space-time object would be useful for many meteo/climatic applications.

Filtering missing pixels in MODIS LST 8-day images through the use of spatial splines also needs to be implemented in the package. Similarly, temporal disaggregation from 8-day images to daily images using splines (in the temporal domain) might be offered as an automated procedure.

Automated mapping using a global model incorporated in the mapping framework is a new approach in this field of mapping. The global model should be iteratively improved with increasing availability (and/or quality) of observations both from ground stations and/or from remote sensing data. Therefore, global modelling of processes (modelled with spatio-temporal kriging) could be performed similarly by storing the global model within automated mapping frameworks.

## Acknowledgements

This study is supported by the research projects *TR36035* and *METEO package - methodological/software solution for automated mapping of climatic variables* , funded by the Ministry of Education and Science of the Republic of Serbia

## References

Hengl, T., Heuvelink, G. B., Perčec Tadić, M., and Pebesma, E. J. , 2012, Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images, *Theoretical and Applied Climatology*, 107:265–277

Kilibarda, M. and Bajat, B.,2012, plotgooglemaps: The R-based web-mapping tool for thematic spatial data, *Geomatica*, 66(1):37–49.

Kilibarda, M., T. Hengl, G. B. M. Heuvelink, B. Gräler, E. Pebesma, M. Perčec Tadić, and B. Bajat, 2014, Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution, *Journal of Geophysical Research: Atmospheres*, 119, 2294–2313.

Pebesma, E. J., 2004, Multivariable geostatistics in s: the gstat package, *Computers & Geosciences*, 30(7):683–691.

Pebesma, E., 2012, spacetime: spatio-temporal data in R, *Journal of Statistical Software*, 51:1–30.

R Development Core Team (2012). R: A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria.

# Movement Regularity Analysis using Geo-Located Twitter Data

E.-K. Kim[1], A. M. MacEachren[1]

[1]GeoVISTA Center, Department of Geography, Pennsylvania State University, University Park, PA 16802
Email: {eun-kyeong.kim; maceachren}@psu.edu

## 1. Introduction

Time-geographic approaches to human traveling behavior have traditionally used origin-destination data (e.g. Cascetta and Nguyen 1988) or activity-travel data collected via diaries and other forms of survey (e.g. Bowman et al. 2001). Origin-destination data is spatially coarse. It can be used to model interactions among places but is of limited use in understanding movement. Survey data can be spatially detailed, but surveys are repeated infrequently and sample size is typically small due to data collection expense and the need to find participants willing to provide longitudinal data (Handy 1996, Calabrese et al. 2013).

As an alternative, researchers have begun to consider location-based and mobile technologies as potential sources of travel activity data. In one example, banknote data was used as a proxy for inter-city mobility in the conterminous U.S. by Brockmann et al. (2006). Cell phone data has been used to provide more detail on individual users' movements, with behaviors explored at different scales including: urban (e.g. Gonzalez et al. 2008, Calabrese et al. 2010, Kang et al. 2012), region (e.g. Calabrese et al. 2013), country (e.g. Krings et al. 2009). Additionally, social media data serve as a proxy for global-scale movements (e.g. Hawelka et al. 2014) as well as national or urban scales (e.g. Azmandian et al. 2013).

The ultimate goal is to enhance understanding of geographic variation in travel behavior in the U.S. and to develop methods for leveraging social media to study spatial behavior. To do so, this paper aims at 1) developing and assessing an algorithm for estimating each Twitter user's residential county by leveraging a full year of individual-based geo-located tweets (i.e. geo-tweets), 2) investigating relationships between tweeter characteristics and geo-tweeting behaviors, and 3) characterizing counties by weekly, daily, and hourly aggregated number of active residential/non-residential Twitter users.

## 2. Research Question

We focus on two sets of research questions: (1) about individual mobility patterns and (2) about characteristics of counties based on collective mobility patterns.

1a) Do individual movement patterns relate to tweeter characteristics, including: the number of tweets, followings, and followers?
2a) How many non-residential users visit particular counties on week days and on the weekend respectively?
2b) What counties are similar to each other?
2c) How does aggregate tweeting behavior vary geographically across the U.S.?

## 3. Methods

### 3.1 Data sets

The analysis reported here is based upon Twitter data collected with the Twitter API from 10/01/2012 to 09/30/2013 throughout the U.S. (Figure 1). After cleaning to remove

duplicates and other data errors, the data set contains approximately 698 million geo-located tweets posted by about 5.5 million users, averaging about 126 each.

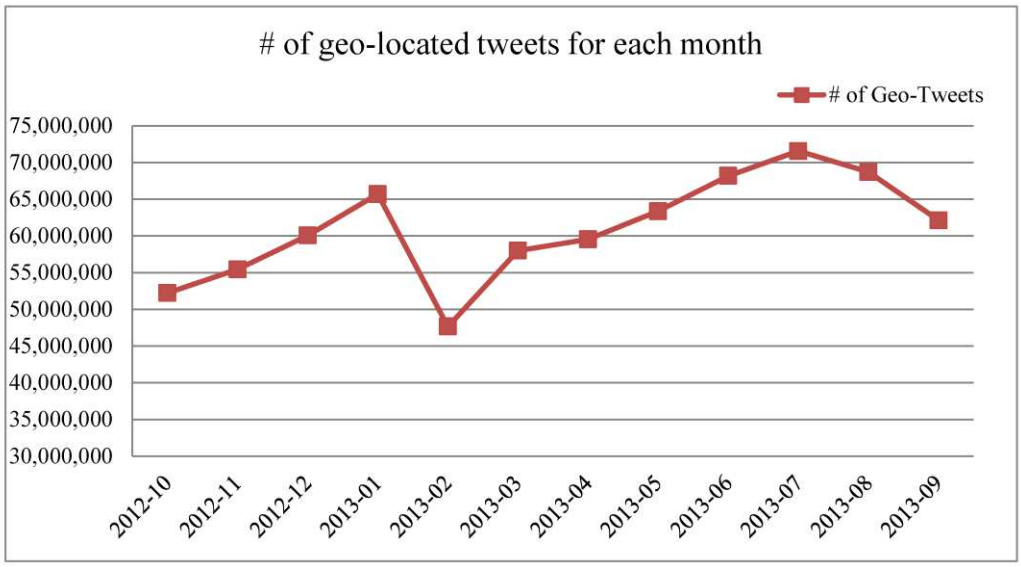



Figure 1. Monthly statistics of geo-located Twitter data used for analyses

### 3.2 Estimation of residential county based the regularity of tweeting behaviors

Geo-tagged twitter data locates individual tweets, thus each tweet provides only the instantaneous location of the tweeter, not the residential or other location. Here, we present a method for estimating and assigning a residential county from a year of geo-tweets. The algorithm used to assign tweeters examines all tweets by an individual for the year, then assigns the individual to the county in which they most regularly posted geo-tweets. Our measurement of regularity is based on frequency of geo-tweets (on days with at least one geo-tweet) by hours, days, and weeks combined with information on mode of geo-tweets and user's time zone in the case of a tie (Figure 2).

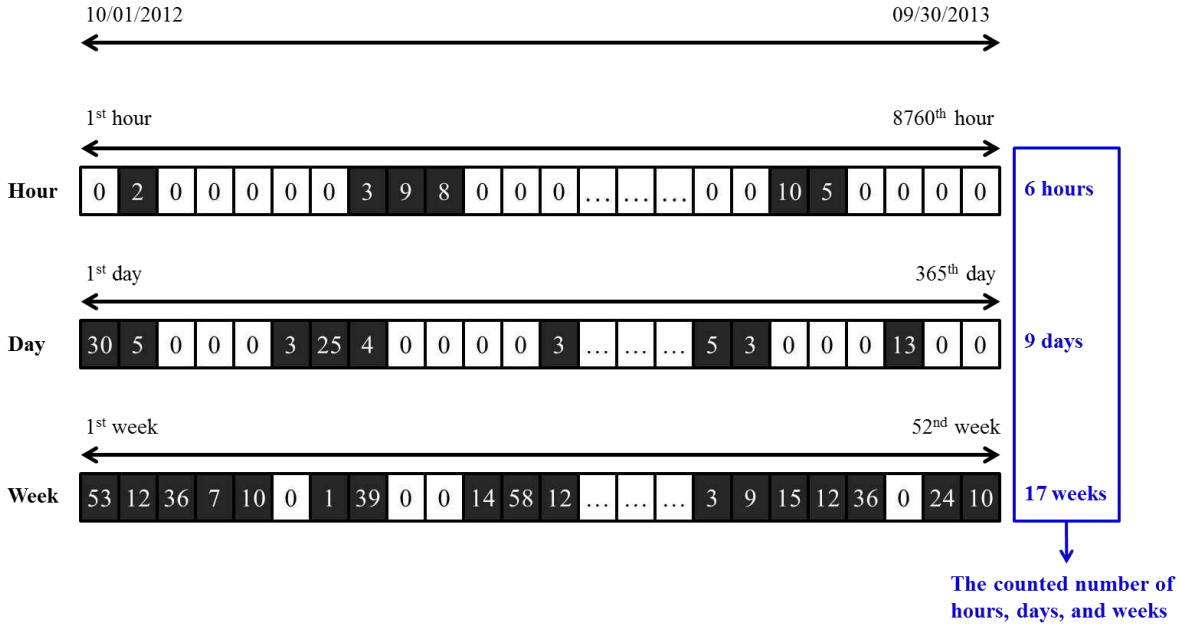



Figure 2. Measurement of the regularity of tweeting behaviors.

### 3.3 Individual mobility and geo-tweeting behaviors

The proportion of hours, days, and weeks over 8760 hours, 365 days, and 52 weeks for which individual users post at least one geo-tagged tweet for each visited county can be measured. The proportion of hours, days, and weeks outside of a residential place can be an indicator of mobility. Exploring the relationship between the proportion of time in the residential county (and other counties) and attributes of tweeting behaviors (e.g. the number of friends, the number of tweets) can be used to uncover how the individual mobility is related to the tendency of online social relationships or Twitter usage patterns.

### 3.4 Characterization of counties through collective geo-tweeting behaviors

Once all users' residential places are estimated, users with the same residential places can be aggregated. This enables looking closely at hourly, daily, and weekly geo-tweeting behaviors of residential users or non-residential users within each county (Table 1 and 2). Some counties have more users on weekends, while other counties on week days. With this weekly pattern, counties can be classified. The number of users for each day can also become a dimension in cluster analysis.

Table 1. An example of the summary of the number of residential users who tweeted in each county (aggregated by day)

| County | Mon | Tue | Wed | Thur | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| Cass, Minnesota | 618 | 615 | 616 | 651 | 673 | 722 | 637 |
| Tarrant, Texas | 96,750 | 96,379 | 97,379 | 99,022 | 100,363 | 102,664 | 102,261 |
| Onondaga, New York | 29,321 | 29,385 | 29,762 | 30,207 | 30,579 | 31,358 | 30,405 |
| Lebanon, Pennsylvania | 4,306 | 4,219 | 4,200 | 4,286 | 4,299 | 4,265 | 4,291 |
| Racine, Wisconsin | 9,262 | 9,320 | 9,270 | 9,409 | 9,469 | 9,463 | 9,651 |
| Franklin, Washington | 2,188 | 2,167 | 2,183 | 2,180 | 2,234 | 2,232 | 2,256 |
| … | … | … | … | … | … | … | … |

Table 2. An example of the summary of the number of non-residential users tweeted in each county (aggregated by day)

| County | Mon | Tue | Wed | Thur | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| Cass, Minnesota | 16,715 | 16,805 | 17,090 | 17,463 | 18,218 | 18,752 | 18,619 |
| Tarrant, Texas | 376,430 | 369,913 | 376,052 | 388,268 | 406,003 | 433,042 | 425,782 |
| Onondaga, New York | 104,628 | 104,401 | 106,017 | 108,095 | 111,886 | 117,622 | 114,772 |
| Lebanon, Pennsylvania | 40,496 | 39,850 | 40,521 | 41,155 | 42,726 | 44,626 | 43,983 |
| Racine, Wisconsin | 52,549 | 52,273 | 53,093 | 53,564 | 55,343 | 57,746 | 57,007 |
| Franklin, Washington | 14,434 | 14,283 | 14,369 | 14,553 | 14,998 | 15,371 | 15,285 |
| … | … | … | … | … | … | … | … |

*Legend 

The smallest two days | The middle three days | The top two days

## 4. Expected Results and Further Work

For individual mobility, correlations between the number of days and weeks outside of a residential place and the number of tweets and friends are expected. As seen in Table 1 and 2, different counties show different daily and weekly patterns.

There is some potential for error in the estimation of residential county (e.g., individuals working a job in a county adjacent to their residential county who only tweet on the job). We plan to assess the assignment of individuals to counties through secondary sources, such as

user profile in those cases where they list meaningful places that are county scale or smaller. Furthermore, this study will classify, map, and analyze spatial patterns of US counties based on temporal patterns of residential/non-residential users' tweeting behaviors and further compare results from our analyses with other reliable movement data including census or tax data.

## Acknowledgements

Data were collected by the Salathe Group at Penn State (http://www.salathegroup.com/) and we appreciate the access to these data.

## References

Azmandian M, Singh K, Gelsey B, Chang YH, and Maheswaran R, 2013, Following human mobility using tweets. In: Zeng LCY, Gorodetsky ALSVI, and Philip S (eds), Agents and Data Mining Interaction, Springer Berlin Heidelberg, 139–149.

Bowman JL and Ben-Akiva ME, 2001, Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A: Policy and Practice, 35(1):1–28.

Brockmann D, Hufnagel L, and Geisel T, 2006, The scaling laws of human travel. Nature, 439(7075):462–465.Calabrese F, Pereira FC, Lorenzo GD, Liu L, and Ratti C, 2010, The geography of taste: analyzing cell-phone mobility and social events. In Pervasive computing, Springer Berlin Heidelberg, 22–37.

Calabrese F, Diao M, Lorenzo GD, Ferreira Jr, J, and Ratti C, 2013, Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation research part C: emerging technologies, 26:301–313.

Cascetta E and Nguyen S, 1988, A unified framework for estimating or updating origin/destination matrices from traffic counts. Transportation Research Part B: Methodological, 22(6):437–455.

Gonzalez MC, Hidalgo CA, and Barabasi A-L, 2008, Understanding individual human mobility patterns. Nature, 453(7196):779–782.

Handy S, 1996, Methodologies for exploring the link between urban form and travel behavior. Transportation Research Part D: Transport and Environment, 1(2):151–165.

Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, and Ratti C, 2014, Geo-located Twitter as the proxy for global mobility patterns. Cartography and geographic information systems, 41(3):260–271.

Joh CH, Arentze T, Hofman F, and Timmermans H, 2002, Activity pattern similarity: a multidimensional sequence alignment method. Transportation Research Part B: Methodological, 36(5):385–403.

Kang C, Ma X, Tong D, and Liu Y, 2012, Intra-urban human mobility patterns: An urban morphology perspective. Physica A: Statistical Mechanics and its Applications, 391(4):1702–1717.

Krings G, Calabrese F, Ratti C, and Blondel VD, 2009, Urban gravity: a model for inter-city telecommunication flows. Journal of Statistical Mechanics: Theory and Experiment, 2009(07):L07003.

Li L, Goodchild MF, and Xu B, 2013, Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography and Geographic Information Science, 40(2):61–77.

# Modelling Communal Tenure using the Social Tenure Domain model

E. Kurwakumire

Tshwane University of Technology, Geomatics Department, Private Bag X680, Pretoria 0001, South Africa
Email: KurwakumireE@tut.ac.za

## 1. Introduction

Land tenure is the relationship between mankind and land according to (Henssen 1995). This is illustrated in Figure 1. This relationship is through rights to land which gives the occupant privileges such as occupancy, ownership and leaseholds. This relationship is well documented in urban areas where freehold and leasehold tenure systems exist as a result of the title and deeds registration systems.

In communal land where communal tenure is predominant, this is rarely the case. Communal tenure is formal in the sense that it is recognised by the statutory law but the actual administration is informal and often characterised by legal pluralism. There are various administrators on the same land to include paramount chiefs, kraal heads, district administrators and rural district councils and as a result, the overall administration is inefficient. With communal land, not only are boundaries for parcels not surveyed and mapped, but the rights to land have also not been documented (Kurwakumire and Chaminama 2012). The closest thing one can get to a land register or cadastre is a list kept by the chief containing the names of occupants within his village without description of the land owned. In other words there is no spatial relationship between occupants and land and this list is always in hard copy format where available.

At the same time, there are various kinds of rights to land to include individual, family and group which gives rise to overlapping rights and claims. Such overlapping rights coupled by multiple administrators all recognised by different statutory instruments brings many problems with actual land and environment administration and in delivery of public services.

## 2. Problem Context

Communal tenure is characterised by a number of challenges. It has a combination of individual, family, group and community property rights. Community rights for example, exist in boreholes, water holes, forests, rivers and pastures. At the same time, there are temporal rights which are seasonal in nature for example, agricultural land has individual or family ownership in the farming season while in the dry season the community has rights of passage through the fields and for their cattle to graze in all the farming fields. As a result, land use is temporal due to the temporal land rights.

There is limited control on livestock numbers and grazing patterns which leads to massive soil degradation and erosion which has a huge impact on the environment. The community can gather pit sand for brick making anywhere suitable. However there is usually not much effort in filling up the pits. These poses a health hazard as the pits and up being breeding grounds for mosquitos in the end which causes malaria. River sand for building is harvested from riverbanks while uncontrolled agricultural activity can in some cases happen along the banks leading to siltation of rivers in the long run. People rely on firewood for cooking and heating which has an impact on deforestation as normally there lacks policies for people to

replant trees to replenish those that have been used. In other words, due to the tenure system, there is limited control of activities happening on the environment.

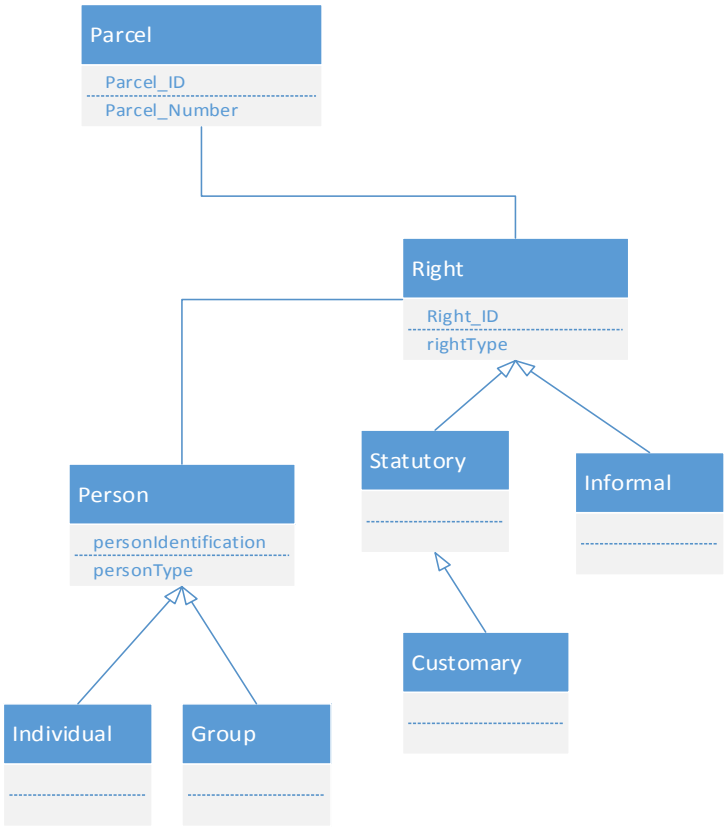



**Figure 1.** Modelling Communal Tenure.

Communal land normally falls under rural land and is normally associated with poverty. One of the contributing factors to poverty in these areas is poor land management and administration. The absence of spatial data deters efficient delivery of public services which are necessary for community upliftment and betterment. Communal land particularly in Africa is the key to future development as urban areas exhaust their land and seek for room to expand.

## 3. Spatial Thinking

In today's world, there is need to combine different spatial data sets so that informed decisions can be made. It is important to know what is where, who is where and even the spread and demographics of the population within a village or ward. This data, referred to as land information, does not exist in communal areas in most cases. At the same time, when the land information exists, it is poor, that is, insufficient, out dated (Mwabujoko 2011) and mostly in hard copy format which is inefficient to share, exchange and distribute. The parcel or property boundaries are not defined, but they are known locally by the occupants. The danger in such a tenure arrangement is the loss of information as generations pass which gives room for larger future land disputes in the future. Such disputes will be difficult to solve as there is no land information to support their resolution.

Communal land has potential for economic activity but it is underutilized because of the nature of the tenure and the subsequent poor administration. It is even difficult to plan future developments as there is no land information available. The information available is mostly

on ward boundaries and village boundaries are often in textual descriptions without actual coordinate information available. As such, they are fuzzy rather than distinct boundaries between villages.

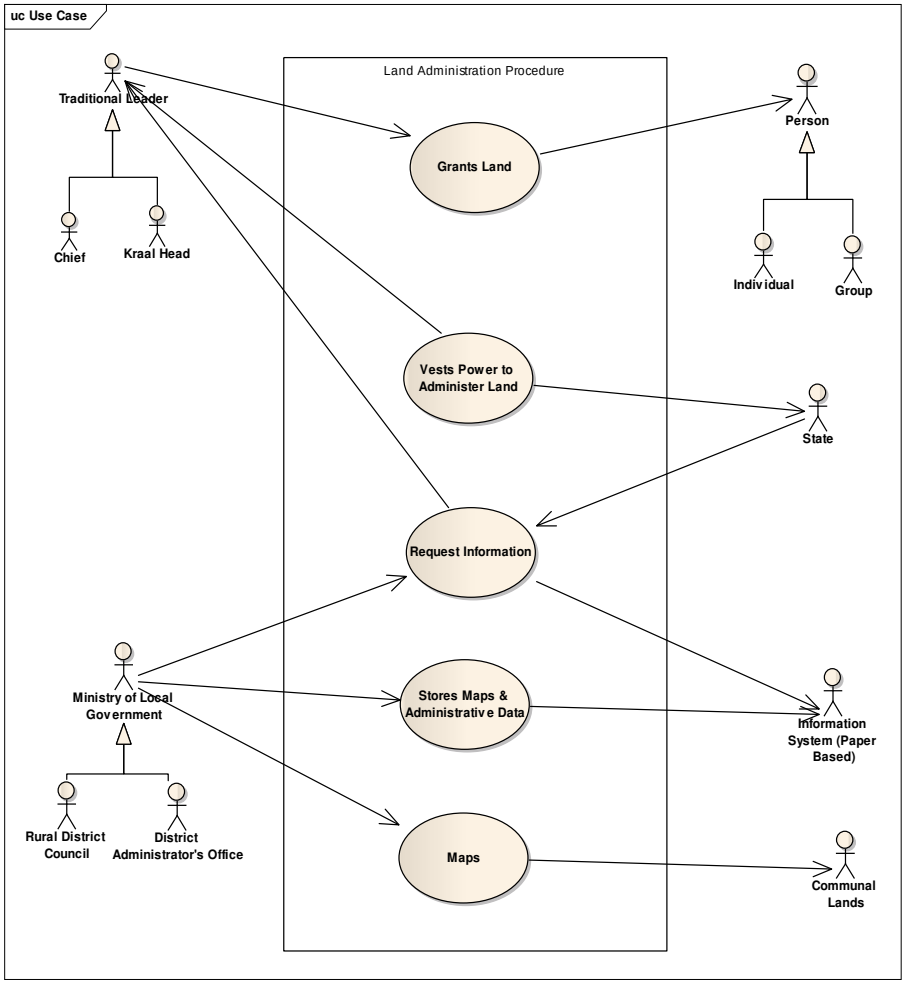



**Figure 2.** Land Administration Procedure.

## 4. Methodology

This study utilises unified modelling language (UML) grounded on the Social Tenure Domain Model (STDM) to model communal tenure and provide a database approach to managing communal tenure (see Figures 1 and 2). The STDM can be used to develop a conceptual schema for social and informal tenure systems Augustinus et al. (2006) in support of pro-poor land administration (Lemmen 2010), societal betterment and good governance in land. The spatial database is developed using largely open source tools as it should be a low cost land information system.

Boundaries are extracted from satellite imagery since in communal land general boundaries are used and can be effectively identified from satellite imagery (see Figure 3). This has been refered to as unconventional approaches to land administration in (van der Molen 2005). Boundary determination should be done using a participatory approach where the community is involved and is free to add additional land information to enrich the data set. Additional information can include names of rivers, mountains, local roads, forests and rivers even year of construction of roads, bridges and boreholes. Communal tenure is largely

grounded on local culture, customs and values which is why chiefs are the local administrators.

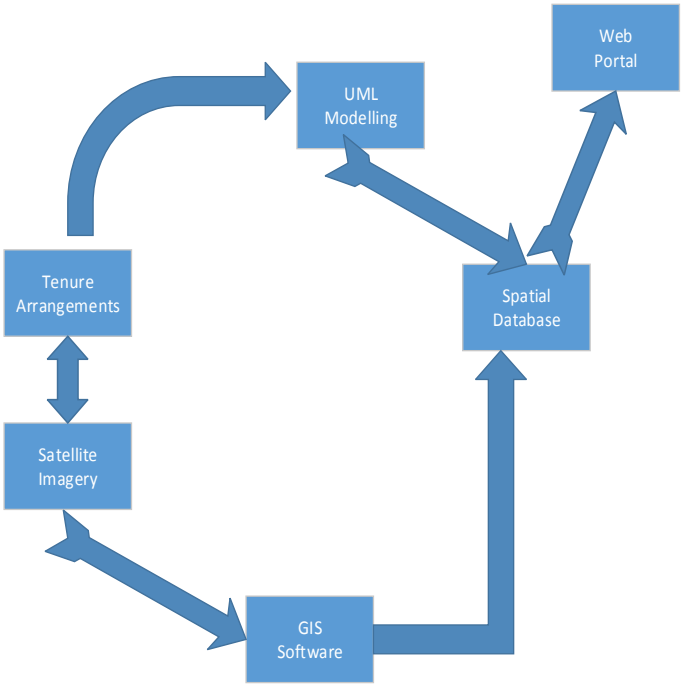



**Figure 3.** Land Information System Design.

## 5. Discussion

The intent of this study is not to change the tenure arrangements as previous studies have shown that land titling is not always the solution. Rather, the modeling which is part of this study is done for two reasons: (1) to model the static situation of communal tenure and (2) to remodel so as to improve it and make the tenure arrangements more flexible while at the same time better securing rights to land.

A database approach towards land administration and management is there to support good governance in land without undermining local institutions while improving public policy formulation and implementation and the location and maintenance of public services.

Fieldwork in support of this study has been done in Lower Gwelo and Goromonzi in Zimbabwe. The end result after the physical design of the database is a community based land information system (CLIS).

## 6. Conclusions

The intend of this study was to develop a mechanism for better managing communal tenure to better administer the land in support of pro poor land management. This would in-turn result in sustainable land and environmental use. The first phase is carried out through mapping the cadastral boundaries through digitising high resolution satellite imagery. The second phase is to link the people, who are the right holders to the parcels, though the associated rights to land. The last phase is to develop a spatial database application and a user

portal that is easy to use. The system should be community centred rather than focused on the technocrats. Within this whole land information system design, education and consultation with the society is key to success.

## References

Augustinus C, Lemmen C, and van Oosterom P, 2006, Social Tenure Domain Model: Requirements from the Perspective of Pro-Poor Land Management. *Proceedings of the 5th FIG Regional Conference: Promoting Land Administration and Good Governance*, Accra, Ghana.

Henssen J, 1995, Basic Principles of the main cadastral Systems in the World. *Proceedings of Modern Cadastres and Cadastral Innovations Seminar*, FIG Commission 7, Melbourne.

Kurwakumire E and Chaminama N, 2012, An Analysis of Data Handling Techniques in Zimbabwe. *Proceedings of the FIG Working Week 2012*, Rome, Italy.

Lemmen C, 2010, The Social Tenure Domain Model: A Pro-Poor Land Tool. In: Uitermark H and Lemmen C (eds), *FIG Publication No 52*, Copenhagen, Denmark.

Mwabujoko EA, 2011, Accessibility of lands information to authorised remote stakeholders in Tanzania, *Article under review for the International Journal of Spatial Data Infrastructures Research,* Submitted February 2011.

van der Molen P, 2005, Unconventional Approaches to Land Administration: A first attempt for an international research agenda. *Proceedings of the ITC Lustrum Conference on Spatial Information for Civil Society*, Enschede, The Netherlands.

# Competing Spatial Optimisation using the k-spatial entropy

Didier G. Leibovici[2], Konstantinos Daras[1] Andy G.D. Turner[1]
[1]University of Nottingham, U.K.
Didier.Leibovici@nottingham.ac.uk
[1]Univerisity of Leeds, U.K.
{K.Daras,A.G.D.Tuner}@leeds.ac.uk

## 1. Introduction

Aggregating geographical data spatially is useful for visualisation, analysis and modelling to support policy formation and decision-making. 2001 and 2011 UK Census Data are available at Output Area (OA), Middle Layer Super Output Area (MSOA) and Ward level. Zoning system optimisation (Daras, 2006, Haynes et al. 2007) becomes useful in optimising the delineation of new aggregating areas to report specific data. Using higher resolution areal units, such as OAs, and optimising heterogeneity or homogeneity of particular descriptors across or within the looked for aggregated zones, similar in scale to MSOAs or Wards, define the problem. Spatial homogeneity of a categorical variable can be measured using the *k*-spatial entropy framework for point data or areal data (Leibovici et al. 2011, Leibovici and Birkin 2014), so that minimum spatial entropy ensures maximum heterogeneity and vice versa. This paper details the following optimisation procedures to aggregate areal units into zones:

- minimum *k*-spatial entropy across the zones and maximum *k*-spatial entropy within each zone (minAmaxW)
- finding a zoning system with maximum *k*-spatial entropy across the zones and minimum *k*-spatial entropy within (maxAminW).

A minAmaxW optimised zoning system will have most homogeneous zones in terms of attribute spatial distribution but with heterogeneous population size, whilst a maxAminW optimied zoning will have regions most similar in total population but with very disparate attribute structuring. Policy-making can potentially use both types. Examples using a microsimulation of the evolution of the population in Leeds between 2001and 2031, are shown. These data are an output from the MoSeS project (Birkin et al. 2009).

## 2. Zoning and entropy

For a set of zones $Z$ aggregating the distribution of a categorical variable $C$ over the sub-zones: a set of proportions $p_{cz}$ with $\sum_{c,z} p_{cz} = 1$ representing the distribution of cases by category and by zone, $p_{cz} = n_{cz}/N$, with $N$ as the total population count, one can use the property of the conditional entropy to get:

$$\begin{aligned} H(C,Z) &=^{def} -\sum_{c,z} p_{cz}\, log(p_{cz}) \\ &= -\sum_z p_{.z}\, log\, p_{.z} - \sum_z p_{.z}(\sum_c p_{c/z}\, log(p_{c/z}) \\ &= H(Z) + H(C/Z) = H(C) + H(Z/C) \end{aligned} \quad (1)$$

with $H(.)$ the Shannon entropy and where $p_{c/z} = p_{cz}/p_{.z}$ with $p_{.z} = \sum_c p_{cz}$ is the conditional probability of the category $c$ from the categorical variable $C$ given the zone $Z = z$. In other words (1) termed the entropy decomposition theorem (Theil 1972, Leibovici and Birkin 2014) insures that the entropy of a categorical variable disaggregated over a zoning is the entropy of the zoning plus the conditional entropy of the variable given the zoning. Moreover one has:

$$0 \leq H(C/Z) \leq H(C) \quad (2)$$

reaching the lower bound when $C$ is completely determined by $Z$ and the upper bound when $C$ and $Z$ are two independent random variables. In regional sciences, a zoning system explaining most of a categorical variable distribution can facilitate policy implementations but working with a zoning system independent of the studied variable facilitates global policy-making expecting to impact equally in each area.

## 3. Self-*k*-spatial entropy

The decomposition in (1) provides a way to communicate with a map (see Figure 1), the variability of categorical data with the local entropies of each zone. Nonetheless the Shannon entropy reflects distributional homogeneity but not spatial homogeneity within each zone. A random permutation of the sub-zones $R$ where $C$ is recorded (the OA here) gives the same entropy. To take into account the spatial pattern, Leibovici (2009) introduced a spatial entropy index based on co-occurrences distributions: the *k*-spatial entropy. A co-occurrence is defined by vicinity, *e.g.,* a maximum distance between *k* occurrences (*k*, the order of co-occurrence, being the number of events to be considered in one collocation). For a given categorical variable the co-occurrence distribution can be seen as multivariate multinomial distribution, *k=3*, giving a tri-variate distribution.

Leibovici (2011) introduced an univariate version, the self-k-spatial entropy (3), easier to understand and compute, looking only at co-occurrences of one category with itself: $p_{c_{oo},d} = p_{iii,d}$ for example with k = 3, so only the hyper-diagonal of the co-occurrence table is used:

$$H^S_{kS}(C,d) =^{def} -{}^{1}/_{log(n_C)} \sum_c p_{cc\,\ldots c,d}\, log(p_{cc\ldots c,d}) \quad (3)$$

The classical entropy is derived from the distribution of the occurrences whilst the self-k-spatial entropy is derived from the spatial co-occurrences for each category. As the self-k-spatial entropy is the normalised Shannon entropy of the co-occurrence distribution, equations (1) with (3) holds with normalising weights coefficients. Nonetheless, in order to make sense of the conditional co-occurrence distribution, the co-occurrences is constrained by the zoning (4), *i.e.,* only co-occurrences whitin a given zone are counted. Cross-boundary co-occurrences for C have to be missed out, so that co-occurrence distributional wise we still have $p(C,Z) = p(C/Z)p(Z) = p(Z/C)p(C)$. Note that Z being a spatial zoning containing the observations with C, the first equation $p(C,Z) = p(C/Z)p(Z)$ is true for both constrained and and unconstrained co-occurrence distributions. Computationally the constrained version of the self-*k*-spatial entropy is faster as parallel evaluations can be done per zone.

$$H^S_{ZkS}(C,d) =^{def} H^S_{kS}(C,\{Z,d\}) =^{def} -{}^{1}/_{log(n_C)} \sum_c p_{cc\,\ldots c,\{Z,d\}}\, log(p_{cc\ldots c,\{Z,d\}}) \quad (4)$$

In Figure 1 the variation obtained from using the co-occurrence distribution rather than the occurrence distribution with the Shannon entropy is seen with the decomposition (top panel) and the conditional entropies or local entropies (bottom panel). The centre of Leeds appears the least homogeneous in relation to social grades but the North-West and South-East areas of the district showing relatively homogeneous Wards.

Can we find a zoning that accentuates this structuring, describing the optimality of the solution in reference to the initial Ward zoning?

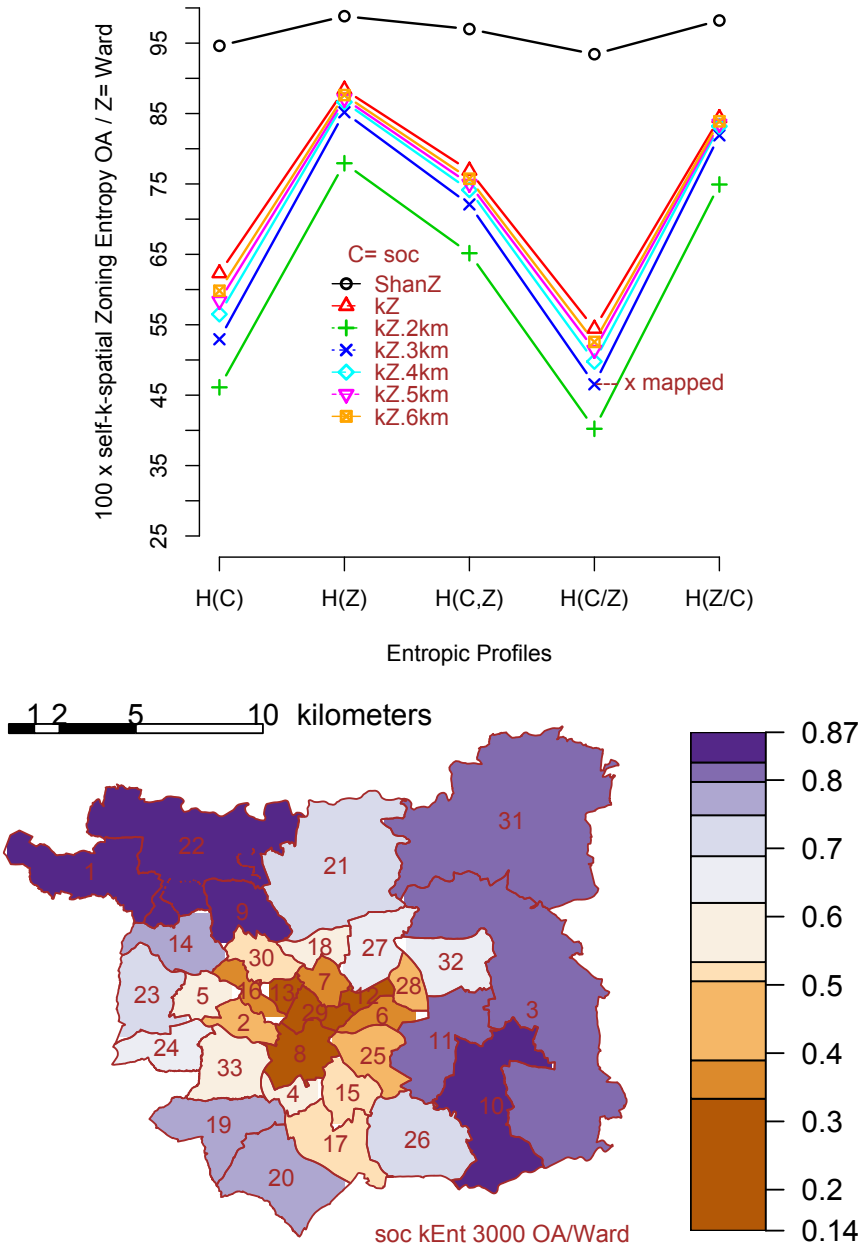



Figure 1: Entropy decompositions for social grades with Ward zoning in 2016 (top panel): Shannon and the constrained self-*k*-spatial zoning entropy, (bottom panel): local values of the within self-*k*-spatial entropies at 3000m.

## 4. Zoning optimisation

A generic homogeneity function for zoning optimisation is the within variance for the grouping in $N_Z$ zones (Daras 2006):

$$Z_{opt} = \underset{\substack{Z \\ Z \succ R \\ Z \in \mathcal{Z}}}{argmin} \; t(y)(Id - P_Z)y \,/(n - N_Z) \qquad (5)$$

where the numerator is just expressing using projectors the sum of squares of residuals from the

local mean for each zone of the attribute $y$, with the zoning $Z$ aggregating the $R$ high resolution zones (noted $Z \succ R$) belonging to a range of valid zoning $\mathcal{Z}$ (defined by a set of constraints such as the compactness of the shapes). The compactness constraint is operating in a competing way during the algorithm and pre-defines the order of testing for local optimum of the objective function.

For categorical variables the zoning often deals only with the proportion of one category (or combined categories). The spatial-entropy index presented in (4) is a good candidate to take into account the whole set of categories to express spatial homogeneity or heterogeneity. Because of the decomposition (1) there is no real competing between maxA and minW in the maxAminW optimisation (6) which also influences the joint entropy:

$$Z_{maxmin} = \underset{\substack{Z \\ Z \succ R \\ Z \in \mathcal{Z}}}{arg\min} \ \alpha / H^{S}_{ZkS}(Z) + \beta H^{S}_{ZkS}(C/Z) \qquad (6)$$

where the optimisation weights : $(\alpha + \beta) = 1$, allow some flexibility along with the set of quality constraints fixed by the ensemble $\mathcal{Z}$ (*e.g.,* number of zones, minimum number of population).

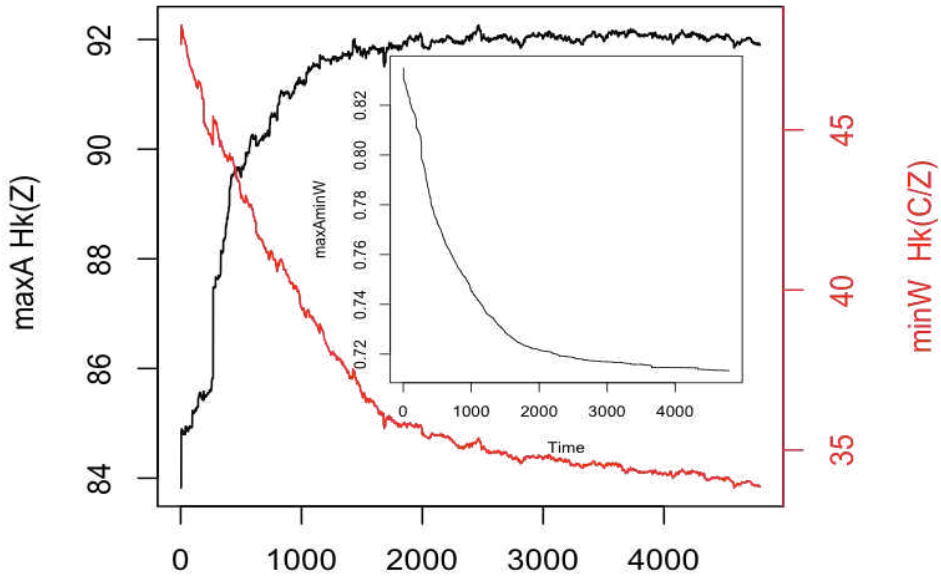



Figure 2:History of the *maxAminW* optimisation (stopping rule: small improvement over last 100 iterations).

As seen in Figure 2, in order to prevent local minima, the algorithm is set to behave alike a simulating annealing optimisation from allowing a little increase of the whole score.

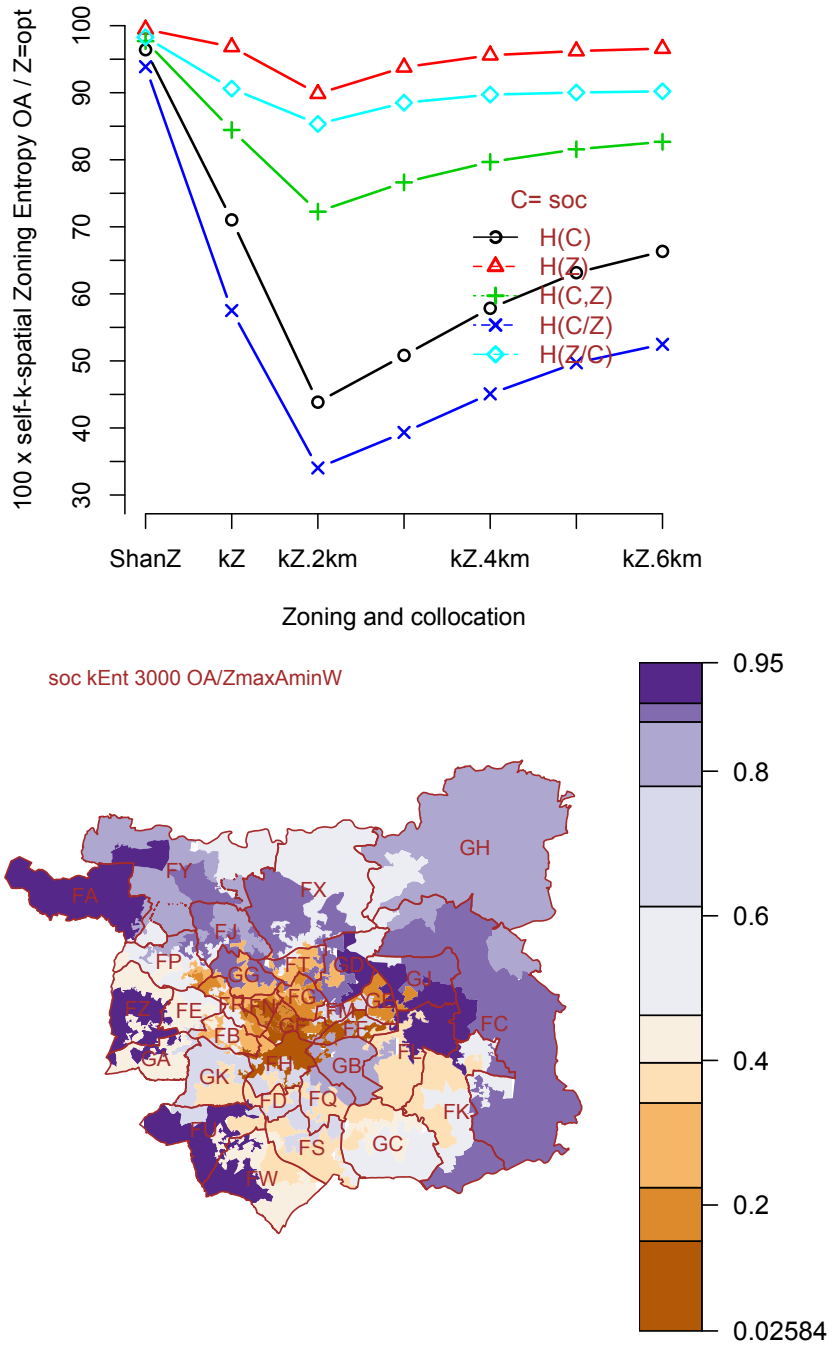



Figure 3: Entropy decompositions with the *maxAminW* optimised zoning (top panel): all for the Shannon and the self-*k*-spatial zoning entropy. (bottom panel ): local values of the within self-*k*-spatial entropies at 3000m (with Wards overlaid).

## 5. Competing compactness

In the optimisation (5) or (6) and the result in Figure 3, the compactness is a competing constraint but it can be included as a real competing optimisation:

$$Z_{maxmin} = \underset{\substack{Z \\ Z>R \\ Z\in Z'}}{arg\min}\ \alpha/H^s_{ZkS}(Z) + \beta H^s_{ZkS}(C/Z) + \gamma Com_{C,R}(Z) \quad (7)$$

where instead of being defined as a rule or threshold for selection as valid zoning in $\mathcal{Z}'$, the compactness is now a third component in the objective function : $(\alpha + \beta + \gamma) = 1$. Because of the differences in variability and in order to insure fair competing, the compactness score has to be normalised against $H^s_{ZkS}(C/Z)$ for its range.

## 5. Discussion

With basic compactness constraint, both optimisation paradigms (*maxAminW* and *minAmaxW*) "converge" quickly and provided useful zonings for this example. The unconstrained value of the self-*k*-spatial entropy at collocation distance 3000m is $H^s_{kS}(C,d) = 0.358$, so quite close to the zoning constrained value $H^s_{ZkS}(C,d) = 0.373$ with the optimal zoning in Figure 3, and better than with the initial zoning of the wards $H^s_{ZkS}(C,d) = 0.529$ in Figure 1. Removing cross-border co-occurrences have here a global "smoothing" effect (for both zonings) which is partially recovered locally within the zoning. The compatible findings reveal the spatial patterns associated with the categorical variable whilst loosing on the re-aggreagated statistic.

For our example the different choices for compactness integration were not crucial for the interpretation but for policy-making and public communication results. The application for zoning optimisation opens up the choices of criteria for this type of spatial clustering where homogeneity and heterogeneity integrates a spatial constraint. The paper proposes using the distribution of co-occurrences and the *k*-spatial entropy framework; the optimality of the solutions are in reference to this and the objective funtion associated to the chosen optimisation paradigm (*maxAminW* or *minAmaxW*).

## 6. Acknowledgements

This work has been funded partially by the ESRC TALISMAN (geosTapial datA anaLysIS and siMulAtioN) project http://www.geotalisman.org

## References

Birkin MH, Townend P, Turner AGD, Wu BM and Xu J, 2009, MoSeS: A Grid-enabled spatial decision support system. *Social Science Computing Review*, 27(4):493-508.

Daras K, 2006, *An information statistics approach to zone design in the geography of health outcomes and provision.* Ph.D. Thesis, University of Newcastle, UK pp 206

Haynes R, Daras K, Reading R and Jones A 2007 Modifiable neighbourhood units, zone design and residents' perceptions. *Health & Place,* (13):812–825.

Leibovici DG, 2009, Defining Spatial Entropy from Multivariate Distributions of Co-Occurrences. *Lecture Notes in Computer Sciences*, (5756/2009): 392-404.

Leibovici DG, Bastin L and Jackson M, 2011, Higher-Order Co-occurrences for Exploratory Point Pattern Analysis and Decision Tree Clustering on Spatial Data. *Computers & Geosciences*, 37(3):382-389.

Leibovici DG and Birkin MH, 2014, On Geocomputational Determinants of Entropic Variations for Urban Dynamic Studies. *Geographical Analysis* (accepted)

Theil H, 1972, *Statistical Decomposition Analysis*. Amsterdam: North Holland.

Wu BM and Birkin MH, 2012, Agent-Based Extensions to a Spatial Microsimulation Model of Demogaphic Change. *In*: Heppenstall et al. (eds) *Agent-Based Models of Grographical Systems, Springer Sciences Business Media*, 347-360.

# A Land Use Classification Method Based on Temporal Spatial Interactions with a Case Study Using Shanghai Taxi Trip Data

X. Liu[1], L. Gong[1], C. Kang[1], Y. Liu[1]

[1] Institution of Remote Sensing and Geographical Information Systems, Peking University, Beijing 100871
Email: {1989liuxi; gongli.mxj; chaoguikang; liuyu}@{gmail.com; gmail.com; pku.edu.cn; urban.pku.edu.cn}

## 1. Introduction

With the rapid development of ICT (Information and Communication Technology), a massive amount of emerging data, such as mobile phone record data, taxi trajectory GPS (Global Positioning System) data and social media check-in data, have been intensively applied for understanding human movements and urban built environments. Land use, a classic issue in geography, also benefits from those data, since they provide opportunities to infer land use via the social function perspective, which is a complement to related remote sensing methods. However, most studies focus on temporally changing activity intensity information (Liu et al. 2012, Toole et al. 2012, Pei et al. 2013) while spatial interaction information is neglected. People's daily travels have strong temporal and spatial patterns (Song et al. 2010), indicating that activity type transition of residents has collective temporal patterns. Since land use closely relates to human activities, interactions between parcels of different land use types also follow different temporal patterns. Decomposing the movement flows according to the interacting land use would provide underlying characteristics for land use classification. In this study, we take advantage of spatial interaction patterns to modify the classification result merely based on temporal activity intensity patterns. Moreover, our method also gives an insight into detecting intensely connected sub-regions of a city.

## 2. Methods

Location and time information of trip origins and destinations, which could be extracted from the ICT generated data, illustrates spatial interactions of a city. To improve land use classification, we decompose the traffic volume of a parcel into traffic volumes between the parcel and other parcels of different land use types. Assuming that the city is discretized into parcels of $k$ land use types, the vector ($k \times 24 \times 2$ dimensions) of parcel $j$ for classification is:

$$X_j = \left[flow_{i,t}^{out} flow_{i,t}^{in}\right] (i = 1, \dots, k; t = 1, \dots, 24) \quad (1)$$

where $f_{i,t}^{out}/f_{i,t}^{in}$ represent trip flows' start/end in parcel $j$ and end/start in parcels of land use type $i$ in the $t$th hour of a day.

Land use of each parcel is needed to determine the vectors. Given that the land uses are unknown and they are also the parameters we want to infer, the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin 1977) combined with unsupervised classification method is applied. EM algorithm begins by assigning land use types randomly to each parcel, and then E-step and M-step are iterated until convergence. In E-step, the vector of each parcel is computed according to the currently assigned land use types; In M-step, the parcels are clustered into $k$ types and the new land use types are assigned to each parcel. We consider the algorithm as having converged when the portion of type-changed parcels between two consecutive iterations is below a certain small value.

Normalization is a key issue to ensure that clustering results correspond to land use types. Two normalization methods showing different facets of spatial interaction are applied and we

call them normalizing "horizontally" (Fig.1 (c)) and "vertically" (Fig.1 (d)). In both methods, inflows and outflows are normalized separately. When normalizing "horizontally", we normalize the 24 temporal flow volumes of each land use type separately as:

$$flow_{it}^{h\_norm} = \frac{flow_{it} - \mu_i}{\sigma_i} \ (i = 1, \dots, k;\ t = 1, \dots, 24) \tag{2}$$

where $\mu_i = \sum_{t=1}^{24} flow_{it}/24$ , $\sigma_i = \sqrt{\sum_{t=1}^{24}(flow_{it} - \mu_i)^2/23}$ . This method means the clustering is based on the temporal volume curves of each land use type that the parcel interacts with. When normalizing "vertically", we normalize the $k$ type volumes of each hour separately as:

$$flow_{it}^{v\_norm} = \frac{flow_{it} - \mu_t}{\sigma_t} \ (i = 1, \dots, k;\ t = 1, \dots, 24) \tag{3}$$

where $\mu_t = \sum_{i=1}^{k} flow_{it}/k$ , $\sigma_t = \sqrt{\sum_{i=1}^{k}(flow_{it} - \mu_t)^2/(k-1)}$ . This method means the clustering is based on the relative volumes of different parcel groups that the parcel interacts with in each hour.

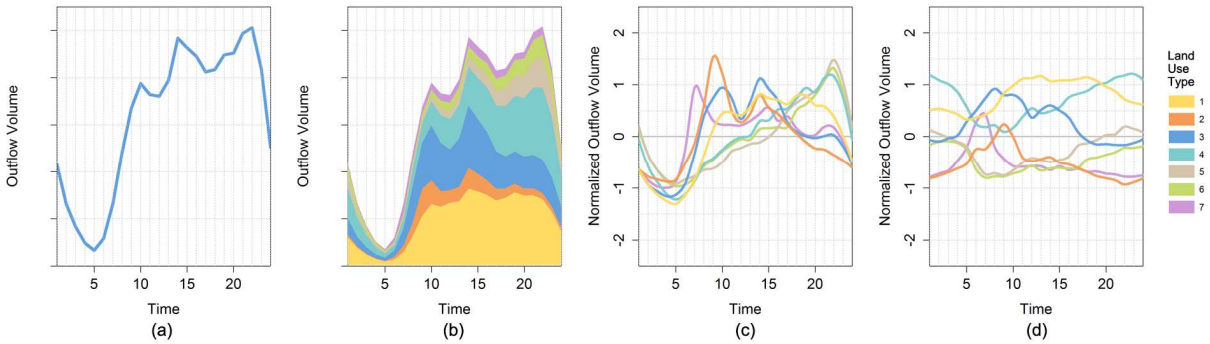



Figure 1: (a) Total trip flows going out from a parcel. Many studies use outflow and inflow volumes to classify land use. (b) The total trip flows originating from the parcel can be decomposed into groups according to land use of destinations. (c) "Horizontal" normalization: the 24 hourly flow volumes of each land use type are separately normalized. (d) "Vertical" normalization: the $k$ land use type flow volumes in each hour are separately normalized.

## 3. Case Study

A case study is conducted in Shanghai using taxi trip data, which are extracted from taxi GPS trajectories from Monday to Thursday in three consecutive weeks (Jun. 1 to Jun. 21, 2009). Each trip record contains the time and location at which a customer is picked up (origin) or dropped off (destination). We discretize Shanghai into 500 $m^2$ grids. K-means clustering method is used in M-steps.

We classify parcels into 7 types and the number of types is determined by interpreting different classification results. In the following part, we refer to classification method based on temporal activity intensity (temporal volumes of customer pick-up and drop-off points in this case) as TAI; temporal spatial interaction with "horizontal" normalization as TSIH; and with "vertical" normalization as TSIV. TAI is considered as the baseline method.

The results of TAI and TSIH are somewhat similar since the vectors of TSIH also contain some information of total temporal activity intensity. The roughly corresponding land uses of each cluster are listed in Table 1 and the grids classified into different groups with the two methods take up 43%. The result of TSIH is less disordered, which better illustrates Tobler's first law of geography, and is more reasonable according to ground truth information from Google Earth (examples in Figure 2 (c), (d)). The normalized activity intensity curves of each cluster of TAI and TSIH are similar, whereas the average flow amounts of each cluster are

completely different (Figure 3). TSIH better classifies parcels with different absolute activity intensities even if the information is not used for classification. These facts all demonstrate that the result of TSIH is more reasonable than the result of TAI. Moreover, by analyzing the interaction patterns between different land uses (Figure 4, 5), we can also better understand human mobility patterns.

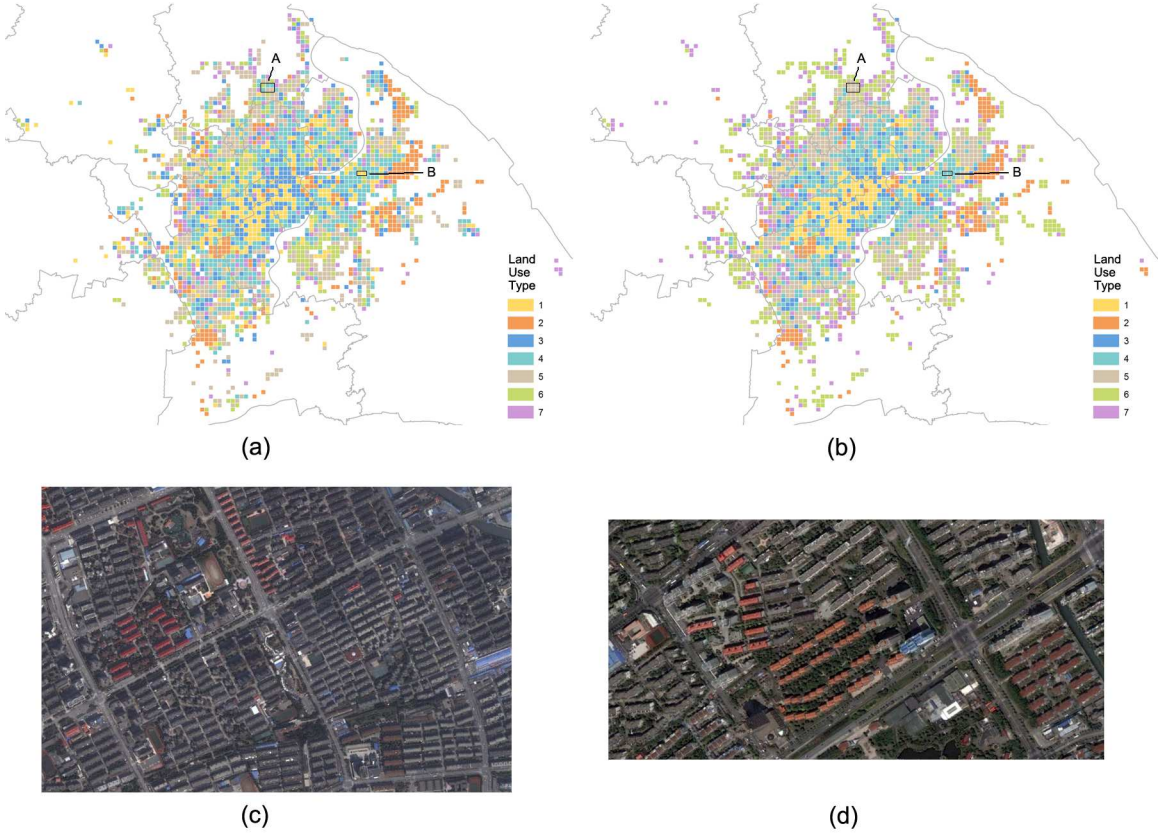



Figure 2: (a) Classification result based on temporal activity intensity (TAI). (b) Classification result based on temporal spatial interaction with "horizontal" normalization (TSIH). As examples, A (c) is a consecutive residential area, but is classified into three land use types with TAI. B (d) is a residential region that is incorrectly classified as an urban commercial area with TAI. Those two regions are both appropriately classified using TSIH.

Table 1. Roughly corresponding land use of each cluster

| Cluster | Land Use |
|---|---|
| 1 | Urban Commercial Area |
| 2 | Business and Industrial Area |
| 3 | Urban Area with Civic Use |
| 4 | Urban Residential Area |
| 5 | Outskirt Urban Residential Area |
| 6 | Suburban Residential Area |
| 7 | Other Land Use Area with Few Taxi Trips |

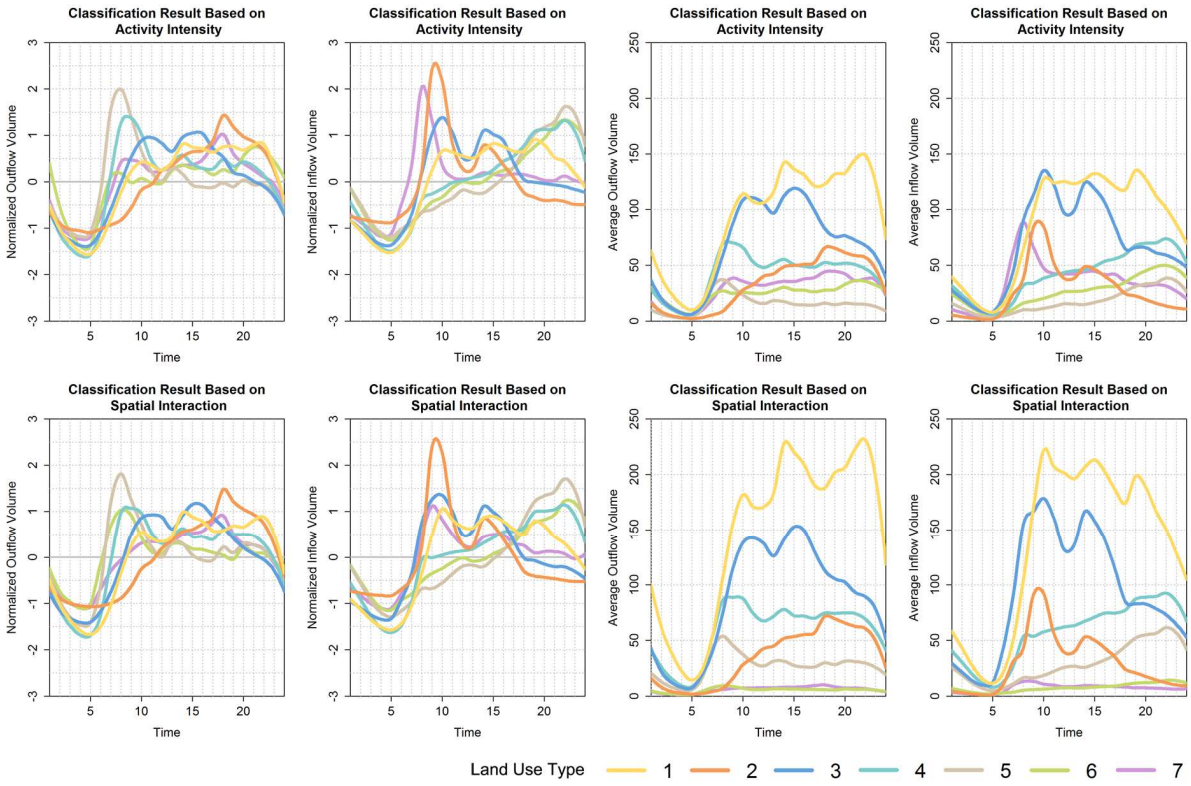



Figure 3: The left figures are the center vector curves of 7 land uses; the right figures are the average flow volumes in a grid of each land use. The upper figures are based on the result of classification using activity intensity; the lower figures are based on the result of "horizontally" normalized spatial interaction.

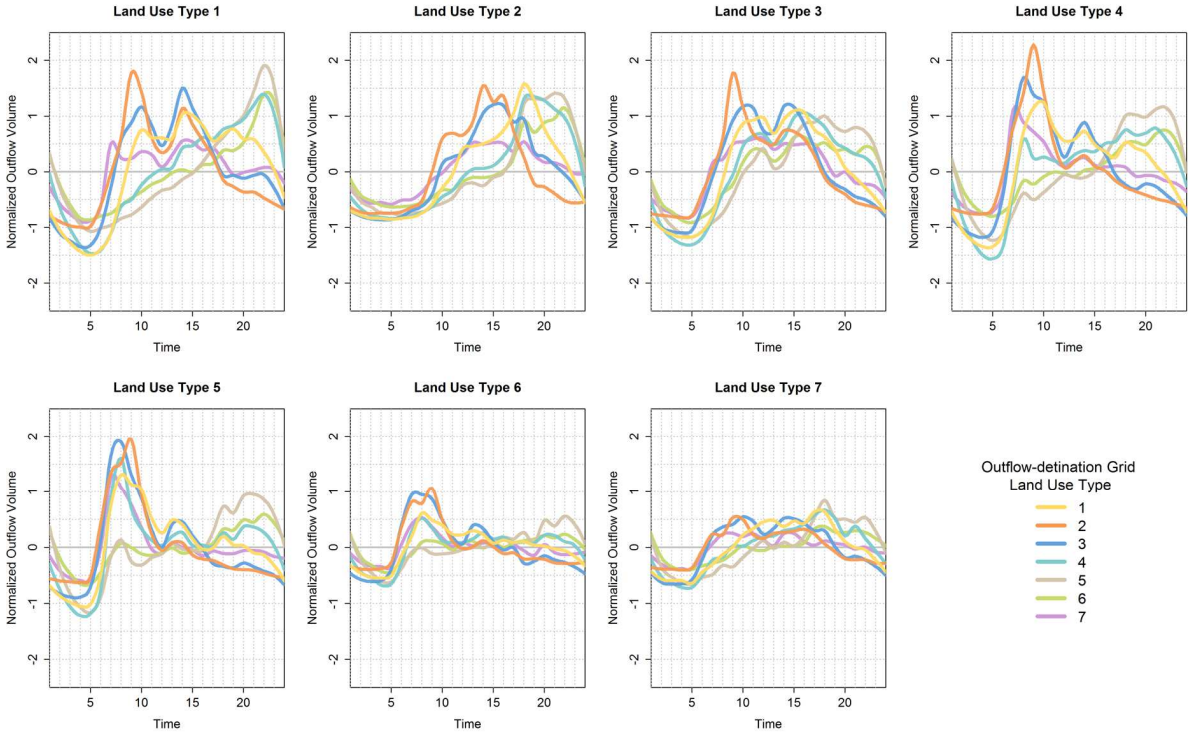



Figure 4: The temporal spatial interactions between land use types from the outflow view.

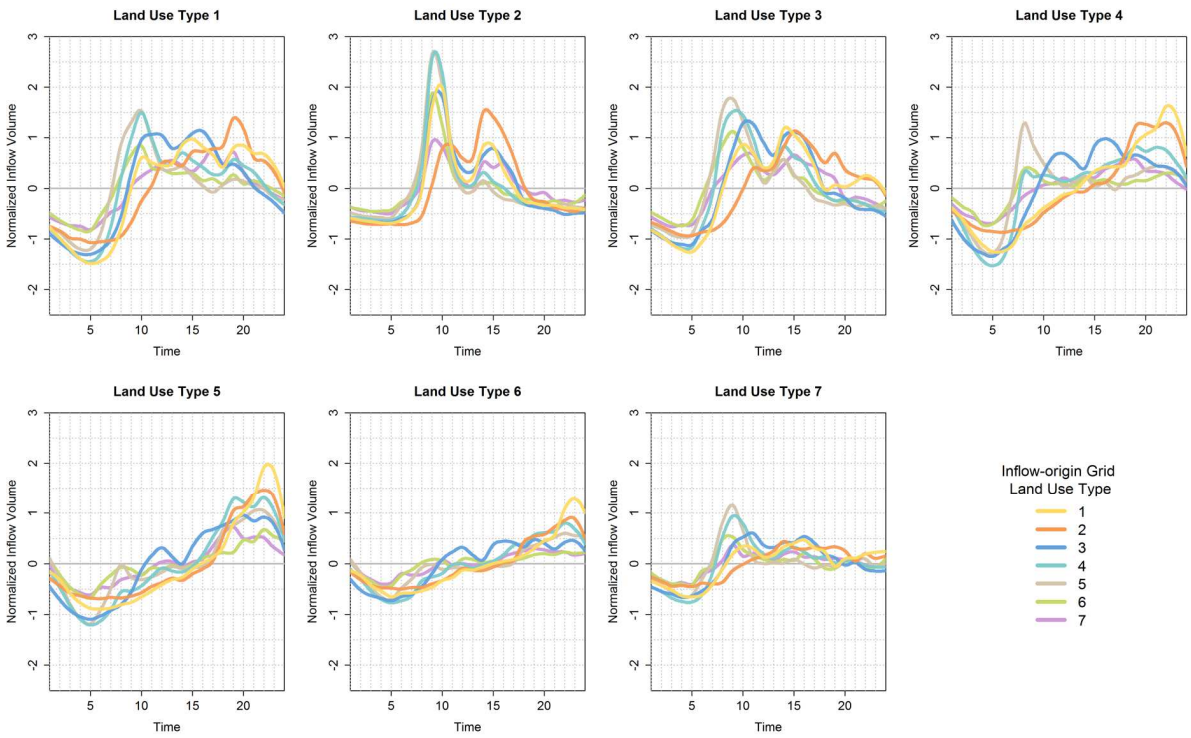



Figure 5: The temporal spatial interactions between land use kinds from the inflow view.

Figure 6 shows the classification result using "vertical" normalization. Surprisingly, the clustering result illustrates spatially continuous sub-regions of Shanghai. The parcels of the same clusters are strongly connected since "vertical" normalization highlights the relative interaction strength between different clusters. This phenomenon may result from distance decay effect: parcels are more likely to interact with nearby places.

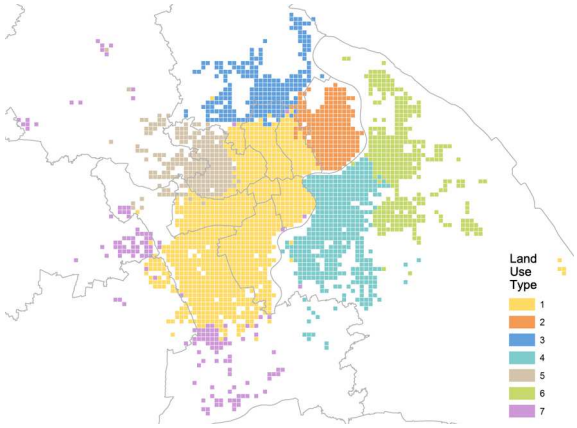



Figure 6: Classification result based on spatial interaction with "vertical" normalization. The result reveals closely connected sub-regions.

## 4. Conclusions and Outlook

This study provides a new land use classification method focusing on spatial interaction, an important perspective neglected in related studies. Spatial interaction information is introduced into land use classification with EM algorithm and proper normalization method. The classification result is improved compared to merely using activity intensity information. Using two normalization methods focusing on different aspects of spatial interaction, interaction similarity and connection intensity are also unified into one framework. Future work would

focus on exploring detailed spatial interaction patterns between different land use types and comparing spatial interaction patterns on weekdays and weekends.

## References

Dempster, A. P., N. M. Laird & D. B. Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal statistical Society, 39, 1-38.

Liu, Y., F. Wang, Y. Xiao & S. Gao (2012b) Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, 106, 73-87.

Pei, T., S. Sobolevsky, C. Ratti, S.-L. Shaw & C. Zhou (2013) A New Insight into Land Use Classification Based on Aggregated Mobile Phone Data. arXiv preprint arXiv:1310.6129.

Song, C., Z. Qu, N. Blumm & A. L. Barabasi (2010) Limits of predictability in human mobility. Science, 327, 1018-21.

Toole, J. L., M. Ulm, M. C. González & D. Bauer. 2012. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, 1-8. ACM.

# Spatial Temporal Analysis of Polygon Movement: Distance and Directional Relationships

J.A. Long[1], C. Robertson[2], T.A Nelson[3]

[1]Centre for GeoInformatics
Department of Geography & Sustainable Development, University of St Andrews
St Andrews, Fife, UK
Email: jed.long@st-andrews.ac.uk

[2]The Spatial Lab
Department Geography & Environmental Studies, Wilfrid Laurier University
Waterloo, Ontario, Canada
Email: crobertson@wlu.ca

[3]Spatial Pattern Analysis & Research Lab
Department of Geography, University of Victoria
Victoria, British Columbia, Canada
Email: trisalyn@uvic.ca

## 1. Introduction

Methods for quantifying spatial relationships in moving polygons continue to be underdeveloped in geographic information science (GISci) despite having numerous application areas (e.g., ecological ranges; Smulders et al., 2012). Methods for polygon movement analysis have lagged behind those for analysing point-based movement trajectories, likely owing to the widespread availability of trajectory data. In fact, there has been borrowing of methods from trajectory analysis with applications better represented using spatial polygons (e.g., hurricanes).

In order to facilitate polygon change analysis, Robertson et al. (2007) developed the STAMP (spatial temporal analysis of moving polygons) framework for quantifying relationships in moving polygons. STAMP facilitates the categorization of polygon movement types (e.g., stable, expansion, contraction) through spatial overlay operations combined with proximity analysis. Robertson et al. (2007) further identified a novel method for measuring directional changes in STAMP polygon categories in order to measure directional changes. The main limitation of the original STAMP implementation is the lack of diversity of methods for computing directional relationships, considering only a single method. Also, distance and shape relationships were ignored in the original STAMP implementation.

The goal of this paper and subsequent analysis is to extend the existing STAMP methodology to include a number of currently available techniques for quantifying distance and directional relationships in moving polygons. We implement the STAMP framework, along with a suite distance and direction indices within the statistical software package R, and we make this package openly available to others.

## 2. Methods

### 2.1 Spatial Temporal Analysis of Moving Polygons – STAMP

The STAMP framework for polygon analysis relies on the overlay between time one (T1) and time two (T2) polygon sets to characterize polygon movement classes (Figure 1). From STAMP overlay analysis polygon relationships are determined from the presence of intersecting polygons, and a distance threshold used to identify disjoint movement events. STAMP classes are arranged in a hierarchy of levels beginning with three classes at the

broadest level (i.e., stable, generation, disappearance) down to eleven rule-based classes representing specific types of polygon movement.

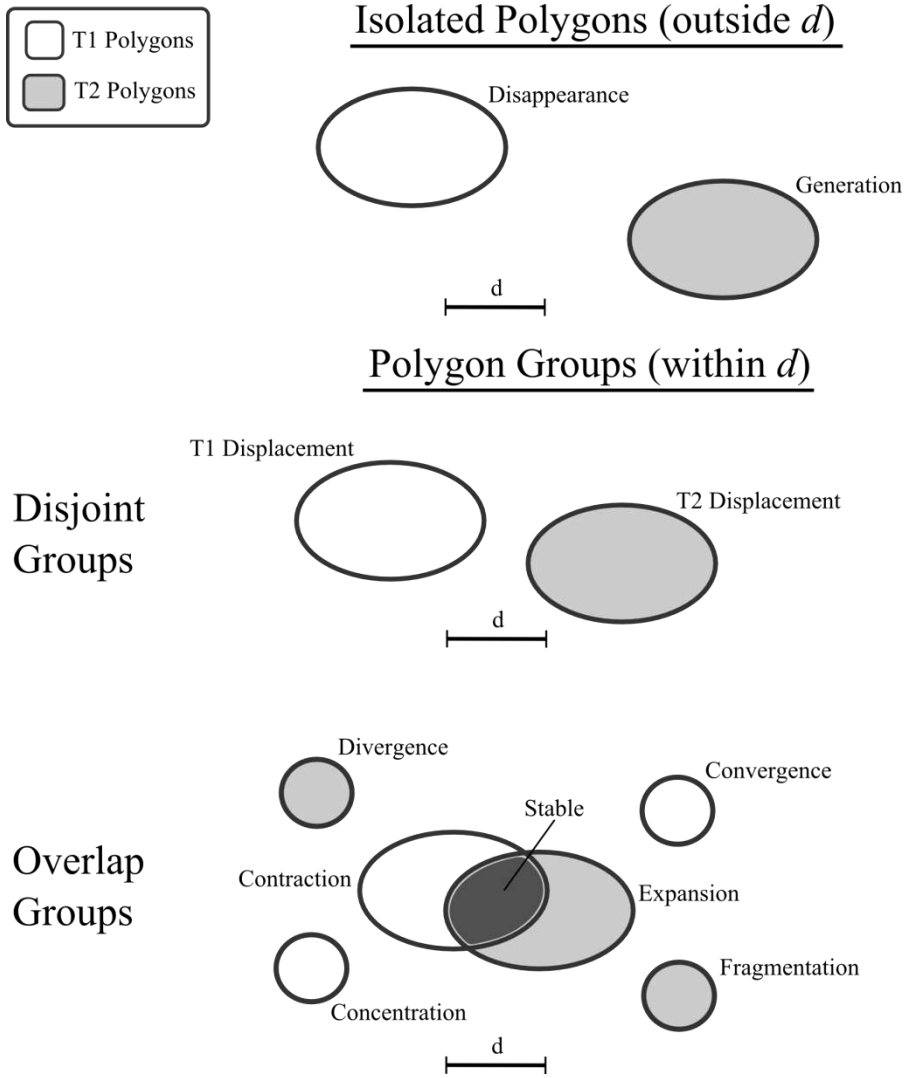



Figure 1: Eleven STAMP classes for various types of polygon movement.

## 2.2 Distance and Direction Indices

There are relatively few distance indices that have been used in examining the distance between polygons. We implement the centroid distance, and Hausdorff distance indices for measuring changes in distance between polygons (Figure 2).

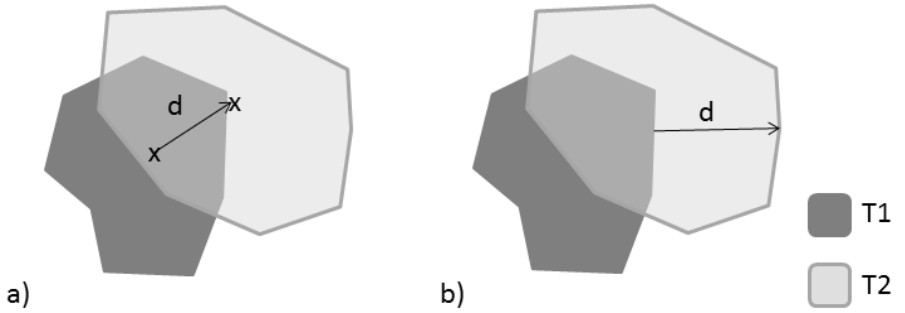



Figure 2: Distance metrics: a) centroid distance, b) Hausdorff distance.

The literature on polygon directional relationships is more diverse, and we examine four methods for exploring directional relationships in polygons. The centroid angle is the simplest, measuring the heading between two polygon centroids (Figure 3a). The cone method (Peuquet & Ci-Xiang, 1987) uses the centroid of the T1 polygons and extends cones outward at specified angles in order to measure the area of T2 polygons contained in each cone (Figure 3b). The modified cone method (Robertson et al., 2007) again uses the T1 centroid, but extends the cones outward to specified locations (e.g., the corners) of the bounding box of the T1 and T2 polygons in order to better account for the shape of the change (Figure 3c). Finally, the minimum bounding rectangle method (Skiadopoulos et al., 2005) relies on the bounding rectangle of the T1 polygon and measures the area of T2 polygons in each of 8 cardinal directions specified from lines extending from the T1 bounding rectangle (Figure 3d).

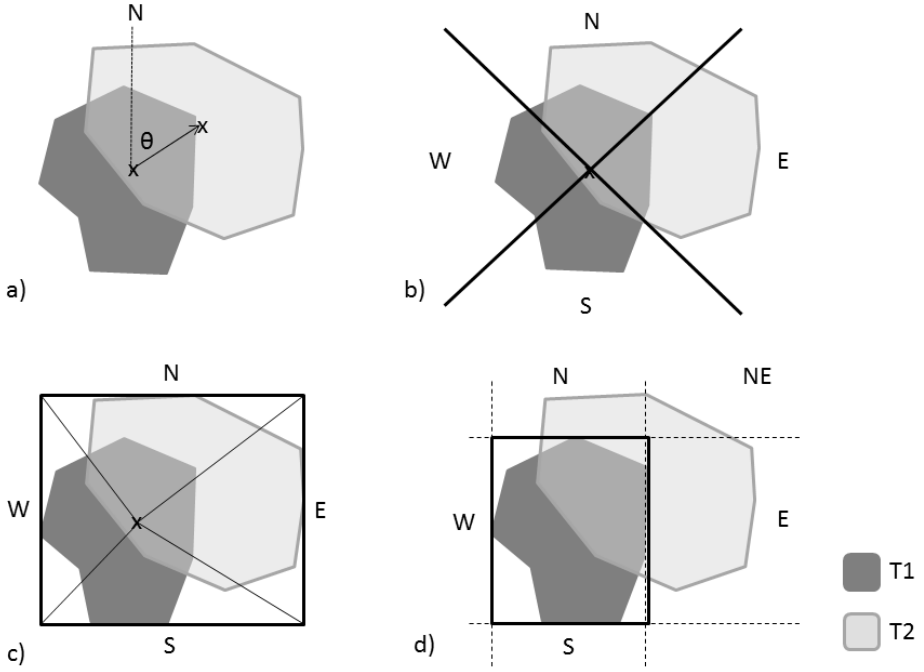



Figure 3: Directional metrics: a) centroid angle, b) cone method, c) modified cone method, d) minimum bounding rectangle method. The centroid angle measure provides a single index of direction, while the other three methods compute the area within directional regions which can be used to investigate the multi-directional aspects of polygon movement.

### 2.3 Case Study: Hurricane Katrina

Here we use a hurricane Katrina dataset to compare and contrast methods for spatial-temporal analysis of moving polygons. Hurricane Katrina data were obtained from the NOAA H*Wind data product (http://www.aoml.noaa.gov/hrd/data_sub/wind.html), which provides a gridded measurement of wind speed covering the hurricane region at 3 hour time steps. Further, we contrast our results from polygon-based analysis with those from trajectory analysis to explore the new inferences afforded through a polygon-based framework.

## 3. Results

We identify subtle differences between distance and directional metrics from polygon-based analysis, and between polygon and trajectory analysis (e.g., Figure 4, Figure 5). Areas when hurricane Katrina was rapidly changing in size are identifiable through STAMP analysis. As well, we can identify some periods where distance and direction are misrepresented by

trajectory-based analysis (i.e., mid-Friday, early Sunday, and late Monday). These regions highlight times when hurricane Katrina was undergoing significant changes either a result of changes in velocity, or through changing directions. Weighted rose diagrams (see poster) are used to display results from the three directional methods employing area-based measurements of direction (i.e., cone method, modified cone method, and minimum bounding rectangle method). Weighted rose diagrams of the area-based measures provide an illuminating graphical representation of the movement of hurricane Katrina through time.

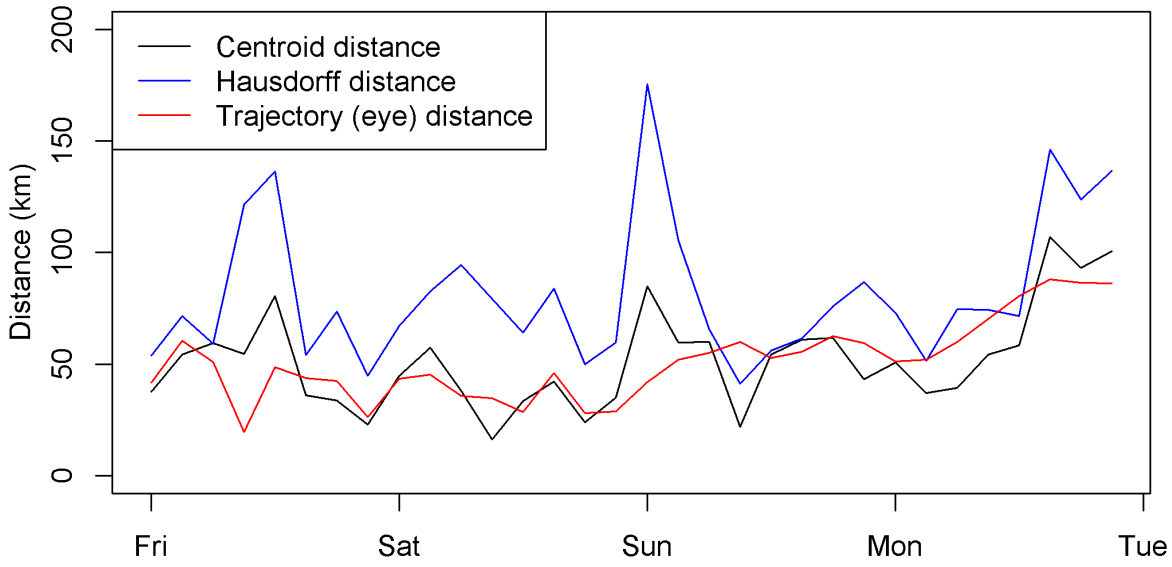



Figure 4: Movement distance of hurricane Katrina using centroid distance, Hausdorff distance, and the trajectory-based measure.

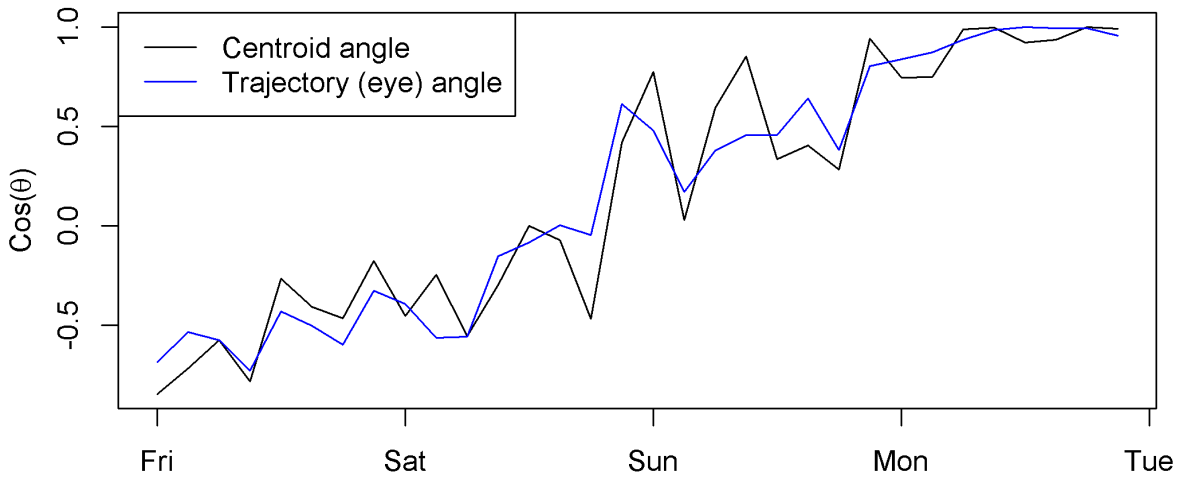



Figure 5: Centroid angle direction measure compared with the trajectory-based directional measure.

## 4. Discussion

The STAMP framework provides a more informative investigation of moving-polygons than using more simplistic trajectory-based indices. Specifically, we find that the Hausdorff distance provided the most useful polygon distance index, identifying more rapid and sudden changes in hurricane Katrina not easily identified through the centroid distance or trajectory-based analysis. As well, the area-based measures of direction (see poster) provide much more

information on directional relationships of moving polygons, and are preferred to single-value direction indices.

We identify several avenues for future research in polygon movement. Further exploration of other metrics for distance and direction relationships is warranted. We see extensions of the Hausdorff metric, for exploring the leading and trailing edges of moving polygons, as well as extracting the Hausdorff movement vector. Novel methods for shape matching of polygons (e.g., Veltkamp, 2001) may also be useful for exploring changes in the shape characteristics of polygons. Similarly, we are exploring shape indices taken from the spatial ecology literature, to simultaneously track polygon movement indices alongside shape indices.

New polygon-based movement analysis will be facilitated through the development of of the 'stampr' package in the R statistical computing environment. The developed 'stampr' package provides a useful platform for future integration of new methods and models.

## 5. Conclusions

Our findings have highlighted the value of polygon-based movement analysis in application areas where appropriate. Further, this work represents a reference point for other researchers wishing to perform polygon movement analysis. To support future research of moving polygons we have implemented the suite of methods presented herein into the 'stampr' package freely and openly available in the statistical software R.

## References

Peuquet, D. J., & Ci-Xiang, Z. (1987). An algorithm to determine the directional relationship between arbitrarily-shaped polygons in the plane. Pattern Recognition, 20, 65–74.

Robertson, C., Nelson, T. A., Boots, B., & Wulder, M. A. (2007). STAMP: spatial–temporal analysis of moving polygons. Journal of Geographical Systems, 9, 207–227.

Skiadopoulos, S., Giannoukos, C., Sarkas, N., Vassiliadis, P., Sellis, T., & Koubarakis, M. (2005). Computing and managing cardinal direction relations. IEEE Transactions on Knowledge and Data Engineering, 17, 1610–1623.

Smulders, M., Nelson, T. A., Jelinski, D. E., Nielsen, S. E., Stenhouse, G. B., & Laberee, K. (2012). Quantifying spatial–temporal patterns in wildlife ranges using STAMP: A grizzly bear example. Applied Geography, 35, 124–131.

Veltkamp, R. C. (2001). Shape matching: similarity measures and algorithms. In SMI 2001 International Conference on Shape Modeling and Applications (pp. 188–197). Genova, Italy: IEEE.

# Modeling Spatio-Temporal Change from Large-Scale High-Dimensional Array Data

Meng Lu

Institute for Geoinformatics, University of Muenster, Germany

Email: meng.lu@uni-muenster.de

## 1. Introduction

This study investigates the feasibility of using array data in large-scale spatio-temporal change modeling through two under-going study cases in climate and land cover change. Statistical approaches for large scale change modeling, such as dimension reduction; time series analysis; time series comparison and classification; joined spatio-temporal analysis will be applied on multi-dimensional array data through multi-dimensional array algebra (Mennis 2009).

Objectives of the study are to develop information extraction technologies with array data by implementing time series analysis, spatio-temporal analysis, and potentially multi-sensor, multi-band, spatio-temporal analysis for change modeling. The research is guided by three questions:

1) How can we meaningfully reduce dimensions spatially and temporally, or thematically?

2) How can we perform complex spatio-temporal statistical computations on multi-dimensional arrays (e.g. extend map algebra to intelligible array algebra)?

3) How can we combine data sets of different dimensionality or different resolutions into arrays?

### 2. Array data

An array is a set of elements which are ordered in discretized space. The number of integers needed to identify a particular position in this space is called the dimension. Each array element is positioned in space through its coordinates. Most natural phenomena can be represented in spatio-temporal arrays once they are quantified in a computer system.

The dimensionality of an array could be unlimited for efficient information extraction and modeling. Examples of most common multi-dimensional arrays include: 1-D time series; 2-D satellite images; 3-D satellite image time series (x/y/t), multi-spectral data (x/y/bands); 4-D spatio-temporal multi-spectral data (x/y/t/bands); subsurface hydrological data (x/y/z/t); and 5-D array with dimensions as different sensors, sensor bands, spatial coordinates, and time (sensors /bands/ x/y/t).

## 2.1 Array data in information extraction

The examples below show the potential of array data in change modeling in terms of spatio-temporal statistics, spectral information extraction, data integration, and parallelized computation:

1. Spatio-temporal statistical analysis: dimension reduction methods, such as Principal Component Analysis (PCA) and its extensions, can effectively extract spatio-temporal variability and shed light on further physical interpretations. As a more complex example, the joined spatio-temporal modeling approach takes into account the spatial dependence of each time point, and thus has the potential to extract more information for noise reduction, movement modeling, and change detection.

2. Multi-spectral information extraction: the spectral variation over time of features on the earth's surface could be represented from the multi-temporal multi-spectral remote sensing image data set (Mello et al. 2013). Compared with selecting several bands for index calculation, this technology brings opportunity in extracting more information for more powerful and accurate analyses.

3. Data integration: remote sensing data are usually either spatially abundant but relatively sparse in time (e.g. satellite imagery) or the reverse (e.g. fixed sensor data). It is possible to integrate data from different sensors and resolutions to produce multi-dimensional array data for more detailed spatio-temporal analysis.

4. Parallelized computation: satellite time series analysis has been well developed and applied to describe historical time series patterns and detect changes in trends and seasonality (Verbesselt et al. 2010ab; Verbesselt et al. 2012; Forkel et al. 2013). Spatio-temporal arrays have the potential to process satellite time series analysis in a parallelized way instead of pixel-wised computation.

## 2.2 Array-based Data Management and Analytics Software System (DMAS)

The programming language R (R Development Core Team, 2008) is able to process and analyze array data. In spatio-temporal statistics analysis mentioned above, for example, the space-time data type in R (Pebesma, 2012) is a step forward in tackling the difficulties in spatio-temporal data processing, analysis, and visualization for integral and joint spatio-temporal approaches. However, R (as well as other data analytics systems) is difficult to scale when dealing with massive data.

The advantages of array-based DMAS are ease of scalability and support for complex queries. Array-based DMAS enables complex computations and statistical analysis with large array data within database management systems. SciDB and Rasdaman are two examples of array based DMAS designed for big data storage, querying, processing and analysis. Both SciDB and Rasdaman partition arrays into sub-arrays, which allows for parallel processing. In addition, SciDB and Rasdaman include features that deal with provenance, uncertainty, versioning, time travel, science-specific operations, and in situ data processing. Rasdaman and SciDB are different in license, system structure, semantics of array operation, and integration with other systems, but they share many similar features and are complementary in their application. For example, the query language of Rasdaman is based on SQL-92 and implements an array algebra through defining a set of operators, while the language of SciDB is a mix of SQL syntax and trees of

algebraic operators. In addition, SciDB provides interfaces to Python and R.

## 3. Methodology and preliminary results

Two study cases were developed to investigate the potential of array data in spatio-temporal change modeling. The two cases focus on different types of array data: fixed-sensor data with regular (e.g. daily rainfall data) or irregular (e.g. rainfall event data) time series; and satellite images. Multi-dimensional array-based database management and analytics systems such as Rasdaman, SciDB, and R will be applied to these cases. In the later stage, study cases might be developed to integrate data coming from different sensors for more detailed information both in space and in time.

The first study case examines the spatio-temporal precipitation variability in two watersheds -- a small semi-arid rangeland in the American Southwest, and the whole Minnesota River Basin in the US. In the American Southwest case, dense rainfall gauges are distributed within the watershed, where convective thunderstorm is the most dominating rainfall type during summer seasons (Goodrich et al. 2008). As initial experiments, PCA has been performed on 30-year precipitation records from the 88 rainfall gauges to detect the spatio-temporal variability of rainfall. The spatio-temporal pattern of weather and climate signals has been successfully extracted and visualized. For a single rainfall gauge, time series analysis has been applied to all rainfall characteristics to describe the trend, seasonality, cyclical patterns and changes in trend and seasonality. In the Minnesota River Basin, increase in water flow has been found during the past decades. Compared with the American Southwestern case the study area of Minnesota River Basin is much larger, the rainfall gauges network is sparser, available records are longer (120 years), and rainfall types are different. Dimension reduction methodologies and spatio-temporal modeling on regular and irregular precipitation array data are explored in this paired comparison case.

The second study case focuses on Juara, Brazil, where the land cover has been changing during the last decade due to deforestation, disturbances and other human activities. The 13 years, 8-day, 250m resolution MODIS satellite images that cover this area has been organized as a dense three dimensional spatio-temporal array and loaded into SciDB for change modeling. The spatial coordinates and temporal information can be retrieved from the array index. As initial experiments, data can be queried and extracted from SciDB to R for analysis and visualization. Spatio-temporal statistics for change detection and monitoring using array data are planned in three methods: 1) perform time series analyses on each pixel in parallel, then summarize time series patterns and changes into spatial pattern; 2) analyze the spatial pattern of each image, then compute time series analyses on the spatial clusters; 3) perform joined spatio-temporal data modeling approach. The results of these three methods are compared to optimize spatio-temporal change detection and real-time monitoring.

The study cases attempt to reduce dimensionality by performing PCA transformation, which results in reduced number of variables that carry useful information. Time series analyses are applied in arrays. For spatio-temporal analysis on multi-dimensional arrays of third or higher degrees, the multi-dimensional array algebra, which considers lags and relationships between neighbors, is developed to enable the spatio-temporal analysis.

# References

Forkel M, Carvalhais N, Verbesselt J, Mahecha M, Neigh C, Reichstein M. 2013. Trend change detection in NDVI time series: effects of inter-annual variability and methodology. Remote Sensing 5(5):2113-2144.

Goodrich DC, Faures JM, Woolhiser DA, Lane LJ, Sorooshian S. 1995. Measurement and analysis of small-scale convective storm rainfall variability. Journal of Hydrology 173(1-4):283-308.

Mello, MP. , Vieira, CAO., Rudorff, B.F.T., Aplin, P., Santos, R.D.C. and Aguiar, D.A., (2013). STARS: A New Method for Multitemporal Remote Sensing. Ieee Transactions on Geoscience and Remote Sensing, 51: 1897-1913.

Mennis J. 2010. Multidimensional map algebra: Design and implementation of a spatio-temporal gis processing language. Transactions in GIS 14, 1: 1-21.

Pebesma, E. 2012. Spacetime: Spatio-Temporal Data in R

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Verbesselt J, Hyndman R, Newnham G, Culvenor D. 2010a. Detecting trend and seasonal changes in satellite image time series. Remote Sensing of Environment 114(1):106-115.

Verbesselt J, Hyndman R, Zeileis A, Culvenor D. 2010b. Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. Remote Sensing of Environment 114(12):2970-2980.

Verbesselt J, Zeileis A, Herold M. 2012. Near real-time disturbance detection using satellite image time series. Remote Sensing of Environment 123:98-108.

# A Crowd-Sourced Taxonomy for the Common-Sense Geographic Domain

David Mark[1], Alexander Klippel[2], Jan Oliver Wallgrün[2]

[1]NCGIA & Department of Geography, University at Buffalo, NY 14261 USA
Email: dmark@buffalo.edu

[2]Department of Geography, The Pennsylvania State University, PA 16802 USA
Email: {klippel,wallgrun}@psu.edu

## 1. Introduction

How are geographic terms, concepts, and their referents related to each other? Is the so-called *geographic domain* a natural partition of reality, or, as some have suggested, is it just an *ad hoc* collection of things that geographers happen to be interested in? These questions are relevant to the various sciences that deal with geographic information.

Taxonomies and ontologies of the commonsense geographic domain were identified as key research goals for Naïve Geography (Egenhofer and Mark 1995) and are central to research on ontologies and semantics in general (eg. Janowicz et al. 2010). In this paper, we present a first step toward a commonsense taxonomy of the geographic domain, derived from a synthesis of behavioral studies of members of the American English-speaking general public.

## 2. Methods

### 2.1 Selection of Terms

Smith and Mark (2001) developed norms for geographic entities by conducting free-listing experiments, asking undergraduate participants to list examples of geographical things; 6 different phrasings of the question were used, and 373 participants provided examples. Participants listed terms for 15 seconds, and provided an average of 5.6 examples each. Together, participants provided 327 different terms. The most frequent example was "mountain" (f=224; 60% of participants).

In the current study, we gave participants 53 terms in a category construction task. These included the 36 most-frequent terms from the Smith and Mark (2001) elicitation norms, plus 17 additional terms with lower frequency, arbitrarily selected from that list. The exact instructions were: "*Please sort on how similar the things are that the words refer to.*"

### 2.2 Participants

The category construction task was administered using CatScan (Klippel et al. 2013) and participants were recruited via Amazon Mechanical Turk (AMT). One hundred participants took part in the experiment (54 female, average age 35.38). Participants received $1 + $.25 for their participation. All participants were native English speakers (we excluded 3 participants with a different language background, and one for indicating his age was 100). All participants live in the US. Mean grouping time was 8 min. Participants created on average 6.68 groups.

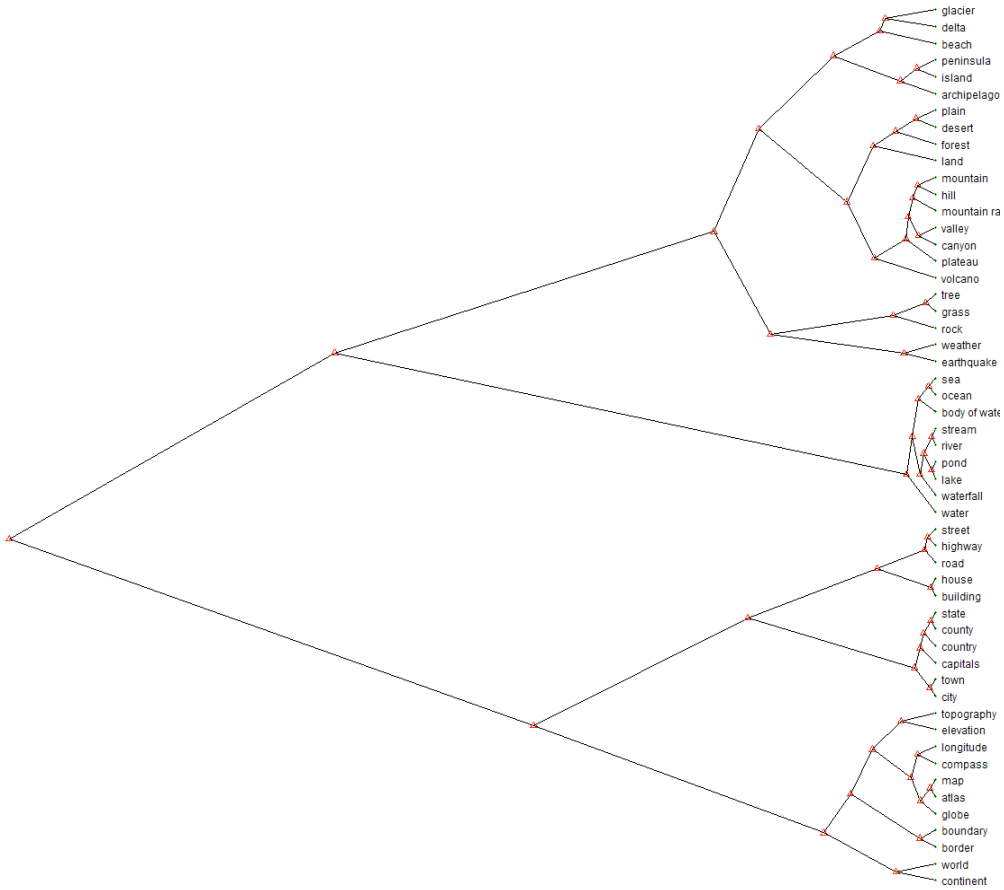



Figure 1. Taxonomy derived from Ward's cluster analysis.

# 3. Results

## 3.1 Interpretation of the Clusters

Figure 1 shows the taxonomy of geographic terms using (exemplarily) Ward's method. It may help our understanding of the dendrogram to use the terms and concepts of taxonomy in biological systematics. One key concept is the *clade*. In biological taxonomy, a *clade* is a group consisting of an ancestor and all its descendants. A *monophyletic* group includes one clade. A *polyphyletic* group includes more than one clade. Of course we realize that this is just an analogy, and that groups of geographic terms do not descend from common ancestral terms. Cluster validation (see section 3.2, below) identified 12 clusters that were robust across all clustering methods. The 8 clades below were considered to be monophyletic if they contain exactly one validated cluster, and polyphyletic if they contain 2 or more validated clusters.

**Eight 'clades'.** By 'cutting' the dendrogram in Figure 1 at an appropriate level, it can be divided into eight clades (sub-trees). For the purpose of structuring the discussion, we refer to the upper five clades as containing relatively natural entity types, and the lower three as containing less natural types (artificial/man-made, abstract concepts). Three of the more natural clades appear to be conceptually homogeneous (which we interpret as analogous to *monophyletic* clades in biology). There is a group that includes all of the water features and only water features. Another

clade is composed of small (sub-geographic) environmental entities (tree, grass, rock), and a third clade contains two types of dynamic entities. We believe that these three clades are uncontroversial. The other two 'natural' clades are heterogeneous (analogous to polyphyletic clades). One of the clades contains two clusters, each of which appears to be homogeneous: seven landform types, and four ecoregion types. The other polyphletic natural clade also contains two validated clusters: three shore-bounded land feature types, and entity types not very similar to any other terms in the study, although they do have a water component: glacier, delta, and beach.

Two of the three clades composed of less 'natural' entity types appear to have conceptual homogeneity. One clade includes components of the *built environment*, and has two coherent sub-clades: street-highway-road, and house-building. A second less-natural clade includes fiat administrative regions (state, county, country) and settlement types (town, city). The remaining clade on the less natural side is deeply polyphyletic, but contains three validated clusters. Four terms denote manipulable artifacts with a geographic purpose: compass, map, atlas, and globe. The placement of the term 'longitude' within this group is strange, but the validated cluster also includes topography and elevation. Another validated cluster within this clade consists of two terms: boundary and border. The last validated cluster within this clade includes world and continent. Some people, including an anonymous reviewer, feel that these terms should be on the 'natural' side of the dendrogram. However, at least in English, 'world' is not a synonym for 'earth', but is more conceptual. 'Continent' also may be thought of as more like a fiat object. But we note that the placement of the world/continent group within the dendrogram is surprising and deserves further scrutiny.

### 3.2 Cluster Validation

To corroborate this finding statistically, we developed a cluster validation technique referred to as cross-method similarity index (CMSI, Wallgrün et al. 2014). Figure 2 shows that the two-cluster solution mentioned above is the most stable conceptualization of the terms used in this experiment. In a nutshell: A CMSI index of 1 indicates perfect correspondence across 100 random samples and three different clustering methods.

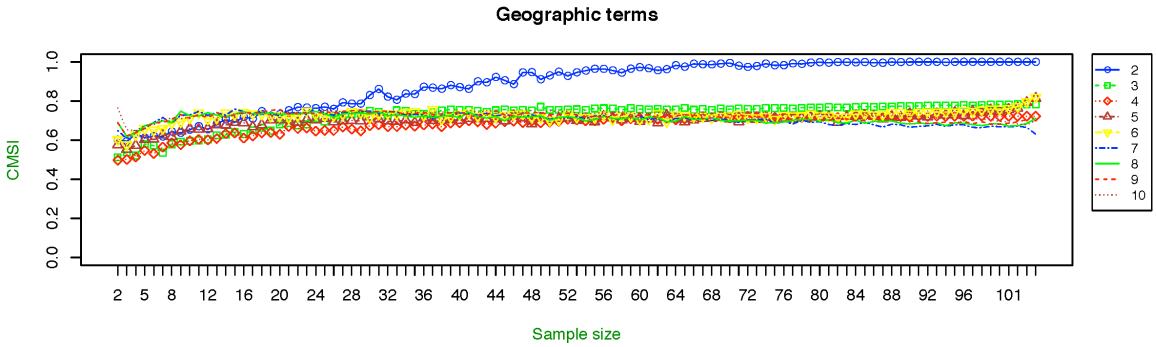



Figure 2. CMSI method for cluster validation. An index of 1 indicates a perfect correspondence across three cluster methods. Each sample-size-value is the average of 100 random samples.

To further understand the conceptual structure (taxonomy) of the terms employed in the experiment, we used an algorithm (Wallgrün et al. 2012) that analyzes tree structure of dendrograms (rather than cluster membership) and compares different clustering methods to validate the results. This approach identified 12 groups of terms that are robust across three

clustering analyses (Ward, average and complete linkage). As noted above, each of these 12 groups has a high degree of internal semantic-coherence, i.e., they appear to make sense.

# 4. Conclusions and Future Work

## 4.1 Artifacts

As noted above, one of the most obvious and clear result is that many participants separated natural geographic entity types from artificial ones. Artifacts apparently present a thorny problem for formal ontologies (cf. Borgo and Vieu 2009), and this presumably will present a challenge for integrating this crowd-sourced taxonomy into formal ontologies.

## 4.2 Future Work

Our next step in this investigation will be the strategic addition of more terms, to confirm or test the over-all conclusions and to fill in semantic gaps. Examples of the kinds of terms we wish to add include: some small non-geographic-related artifacts such as chair, book; some small animals; some very large mobile artifacts such as ships; some constructed water features (will they more often be grouped with water features or with artifacts?); some terms for kinds of wetlands; and some additional land cover types or biomes, such as tundra, prairie, and savanna. Eventually we also wish to test similar sets of terms for other languages.

Another extension of this work would be to code the taxonomy in an ontology-coding framework such as Protegé. This will require the introduction of, and naming of, internal nodes for the taxonomy, ideally from an established Upper-Level Ontology such as DOLCE or BFO.

# 5. Acknowledgements

Comments from Gaurav Sinha and from an anonymous reviewer were useful and are appreciated.

# 6. References

Borgo, S., and Vieu, L., 2009. Artefacts in Formal Ontology. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 23, 3-21.

Egenhofer, M. J., and Mark, D. M., 1995. Naive Geography. In Frank, A. U. and Kuhn, W., (Eds.), *COSIT 1995*, Berlin: Springer, pp. 1-15.

Janowicz, K., Schade, S., Bröring, A., Keßler, C. Patrick Maué, and Christoph Stasch, 2010. Semantic enablement for spatial data infrastructures. *Transactions in GIS*, 14(2), 111-129.

Klippel, A, Wallgrün, J O, Yang, J, Mason, J S, Kim, E-K, Mark, D M, 2013, Fundamental cognitive concepts of space (and time): Using crosslinguistic, crowdsourced data to cognitively calibrate modes of overlap. In Tenbrink, Stell, Galton, Wood (Eds.), *COSIT 2013*. Berlin: Springer, pp. 377–396.

Smith, B., and Mark, D. M., 2001, Geographic categories: An ontological investigation. *International Journal of Geographical Information Science*, 15 (7), 591-612.

Wallgrün, J O, Yang, J, Klippel, A, Dylla, F, 2012, Investigations into the cognitive conceptualization and similarity assessment of spatial scenes. In Xiao, Kwan, Goodchild, Shekhar (eds.) *Geographic Information Science-7th International Conference, GIScience 2012*, pp. 212 – 225.

Wallgrün, J O, Klippel, A, & Mark, D M, A new approach to cluster validation in human studies on (geo)spatial concepts. *GIScience 2014 (extended abstracts).*

# Inter-Organizational Spatial Networks

Mahbubur Meenar, PhD

Temple University, 580 Meetinghouse Rd, Ambler, PA 19002
Email: meenar@temple.edu

## 1. Introduction

This paper presents an exploratory examination of inter-organizational networks (IONs) and partnerships among Philadelphia-based nonprofit organizations (NPOs) that offer food-related programs. Social capital and social networks have spatial or geographic dimensions (Poon et al. 2012), even though sociologists have always focused on only relational space (Metcalf and Paich 2005). In 2005, Michael Batty indicated the spatial aspect of social network analysis (SNA) as one of the next steps for future research (Batty 2005: 168). Yet today, few studies have tried to integrate ION research and GIS based SNA. GIS has been integrated to digital social network analysis (d-SNA) (i.e., Facebook or Twitter networks), but integration of relational and spatial network analysis (SPNA) is still at the infancy stage.

Understanding the geographic distribution of NPOs and their programs can be a useful starting point to realize community resources and their geographic proximity to local needs related to community social issues such as food insecurity. The results and findings of this research would help NPOs make better decisions in implementing their activities and programs. In addition, this study is expected to advance the new line of research that focuses on the integration of GIS and SNA.

## 2. Context, Methodology, and Data

This study was based on the City of Philadelphia. One in four residents in this city are at risk for hunger, more than double the rates reported at both the national and state levels. In many food insecure neighborhoods, disadvantaged residents do not have easy access to healthy and fresh food, have poor food habits, and have diet-related chronic health conditions. The city, on the other hand, is nationally known for many of its NPO-driven initiatives and partnerships.

Since there are no universally accepted neighborhood boundaries available in Philadelphia, this study used planning district boundaries (n=18) as geographic units of analysis. Overall, 153 NPOs were studied and their primary office locations were geocoded. The study methodology used GIS-based SPNA and SNA (i.e., network density, bridging and bonding network, etc.). Primary data were collected from a 36-question online survey (with 79% response rate) and interviews of 38 NPO representatives, and from online sources, i.e., websites and social networking sites.

## 3. Analysis, Results, and Interpretation

### 3.1 Inter-Organizational Partnerships

The majority (81%) of NPO representatives reported that they received funding, such as direct funds, transfer of funds, and sub-contracts from other NPOs. The same percentage of NPOs partnered with others to execute a program or policy. About 67% prepared grant proposals in collaboration with other NPOs and 28% provided funding. About one-third of all NPOs surveyed (n=38) did not report any partners. These are primarily small-scale NPOs or grassroots organizations. Few NPOs make only short-term financial partnerships with others.

These partnerships often are manifested in the form of donations and tools or volunteer exchanges.

### 3.2 Spatial Network Analysis

Figure 1 showcases IONs of NPOs that were included in this study. NPOs are spatially concentrated toward the central part of the city, so this area has a higher presence of network connection lines. The 38 NPOs without any partners are presented as single points. The ION is spread throughout a portion of the whole city, not concentrated in some smaller "network neighborhoods", as illustrated by Hipp et al. (2012).

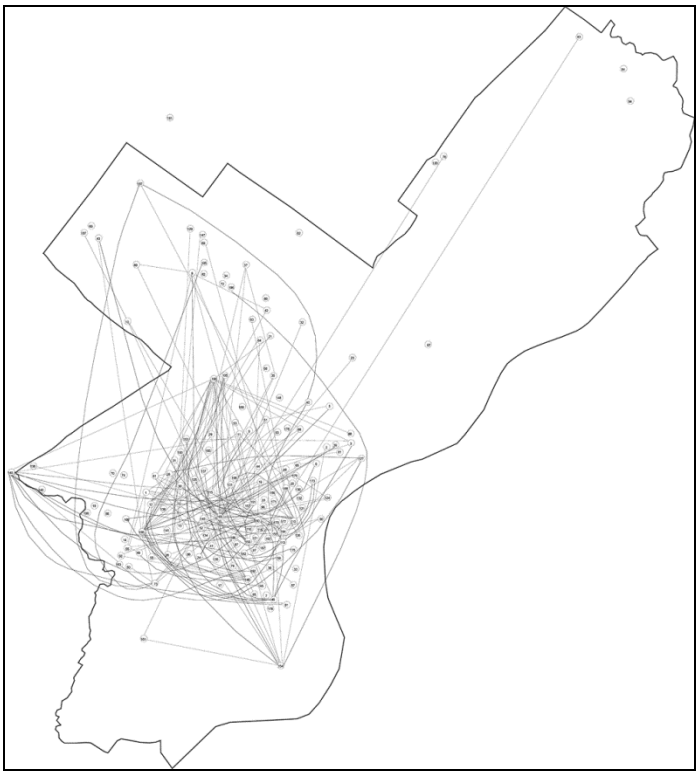


Figure 1. Organizational network of food-related NPOs with other NPOs with similar agenda.

### 3.3 Density of Network

Density of network refers to the total number of connections between organizations in a network divided by the total number of possible connections. An ideal, fully-connected network is supposed to have a density of 1.00. Density of network was calculated for each NPO, resulting values from 0 to 0.0064. Then the average values were calculated at the planning district level (see Table 1). Although Figure 1 provided an impression of a dense organizational network in Philadelphia, actual density values of all these NPOs in planning districts were very low and sometimes spatially misplaced.

Table 1. Food-focused NPOs and their density of network

| Planning District | # NPOs | Average Density of Network |
|---|---|---|
| Central | 52 | 0.0311 |
| Central Northeast | 2 | 0.0002 |
| Lower Far Northeast | 0 | 0.0000 |
| Lower North | 15 | 0.0052 |
| Lower Northeast | 1 | 0.0002 |

| | | |
|---|---|---|
| Lower Northwest | 5 | 0.0011 |
| Lower South | 1 | 0.0015 |
| Lower Southwest | 1 | 0.0004 |
| North | 7 | 0.0046 |
| North Delaware | 0 | 0.0000 |
| River Wards | 5 | 0.0017 |
| South | 9 | 0.0051 |
| University/Southwest | 23 | 0.0131 |
| Upper Far Northeast | 5 | 0.0003 |
| Upper North | 2 | 0.0004 |
| Upper Northwest | 15 | 0.0035 |
| West | 6 | 0.0018 |
| West Park | 4 | 0.0009 |

### 3.4 Spatial Bonding and Bridging Network

Geographic or spatial bonding and bridging networks of all the 153 NPOs were tested. The understanding of these two types of networks followed literature on bonding and bridging social capital or networks with or without closure (Putnam 2001). When an NPO was connected to another NPO in the same planning district, the network was termed as a bonding network and when connected to an NPO of another district, the network was termed as a bridging network. The calculation was done for each NPO located in each planning district.

The results from this analysis demonstrate NPOs generally have more bridging partners than bonding. 85 NPOs have zero bonding partners, 43 NPOs have only one bonding partner each, 15 NPOs have two partners each, and the remaining 10 NPOs have three or more bonding partners. All of these NPOs with three or more bonding partners are located in the Central District. On the other hand, 54 NPOs have zero bridging partners, 41 NPOs have only one bridging partner each, 16 NPOs have two partners each, and the remaining 42 NPOs have three or more bridging partners. NPOs with higher numbers of bridging partners are also located in the Central District. Most interviewees considered these NPOs as key or central players in Philadelphia's food systems network. From this analysis, it can be interpreted that the more bridging networks an NPO has, the more central it is to the whole organizational network. This interpretation is consistent with Kropczynski and Nah (2010). Figures 2 and 3 provide two examples.

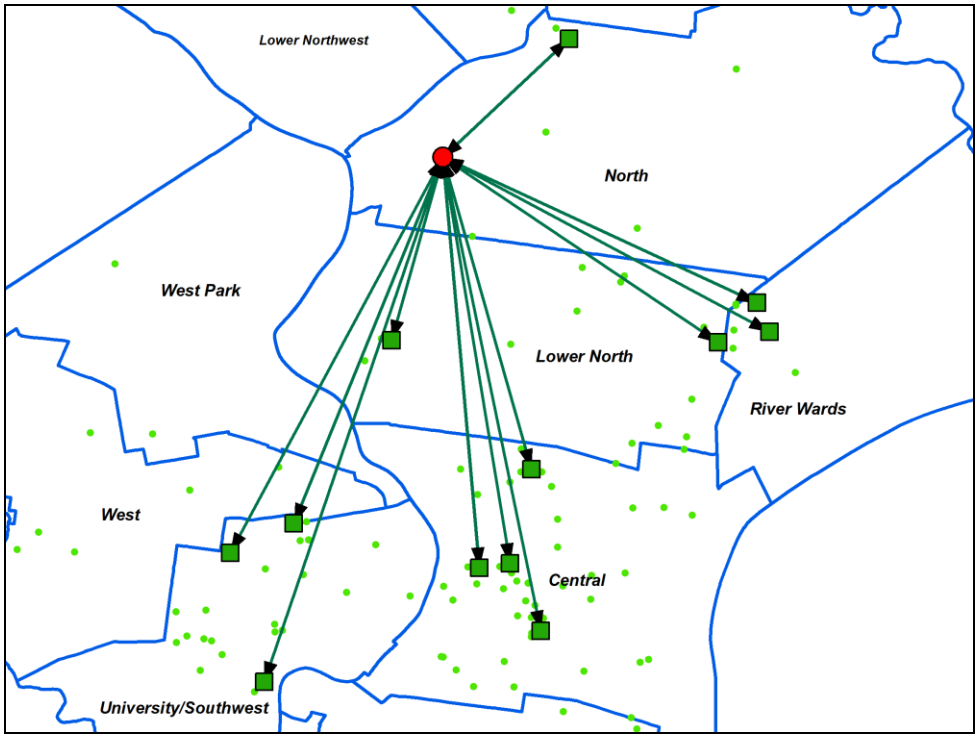



Figure 2. SHARE Food Program (circle) and its partners (squares)

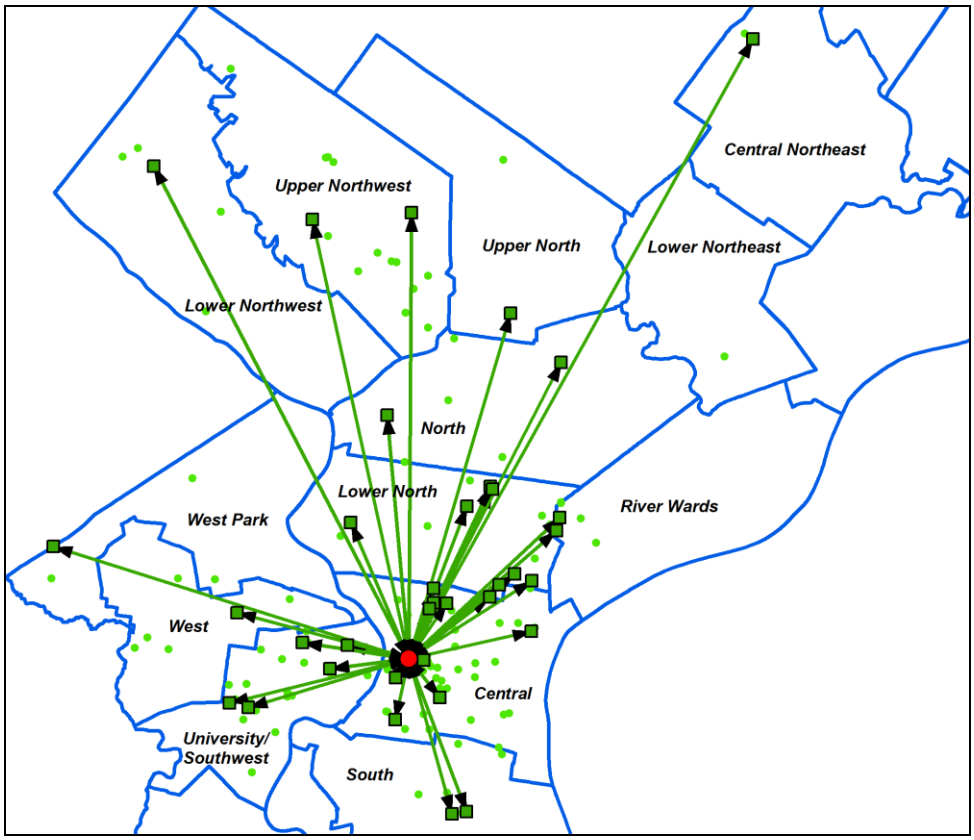



Figure 3. Pennsylvania Horticulture Soceity – PHS (circle) and its partners (squares)

## 4. Discussions and Conclusion

According to this study, geographic boundaries did not directly influence organizational networks. Since most larger and issue-based NPOs were located in the Central District, many

place-based NPOs rooted in different neighborhoods were connected with them, regardless of their distances or geographic boundaries. In terms of partnerships, most NPOs preferred common interests or agendas, financial standing, and political connections over geographic proximity. This finding was consistent with Chen and Graddy (2010).

Unlike what Strauss (2010) suggested, NPOs studied in this research formed more bridging partnerships than bonding, geographically speaking. Although NPOs within the same neighborhoods always compete with one another to catch a funder's attention, there is no alternative to strengthening coordination and partnerships not only among NPOs, but also with other organizations such as governments and institutions. Coordination efforts among NPOs and smaller agencies (such as food cupboards or grassroots community gardens) can be made stronger at both the local and state levels.

The intellectual merit of this research is both theoretical and methodological. GIS has been integrated to social media analysis, but integration of SNA and GIS is still at the phase of infancy. This study joins with the argument that there is a lack of integration between the scientific community of ION, SNA, and SPNA research. GIS-based tools are not yet easily available for conducting SPNA research, but this study has tried to experiment with the theme – a combination field that has begun to be explored only recently. According to this research, SPNA calculation is possible using a variety of tools available in Desktop GIS software, but the tools are disjointed. In addition, the calculations for network density and bridging or bonding network were done manually. It is important to merge the tools available in typical SNA software with desktop GIS, so that regular users, including NPOs, can utilize these tools and run their own ION analysis.

## References

Poon J, Thai D and Naybor D, 2012, Social capital and female entrepreneurship in rural regions: Evidence from Vietnam. *Applied Geography*, 35: 308-315.

Metcalf S and Paich M, 2005, Spatial dynamics of social network evolution. *23rd International Conference of the System Dynamics Society*, Boston, USA.

Batty M, 2005, Network geography: Relations, interactions, scaling and spatial processes in GIS. In: Fisher P and Unwin D (eds), *Re-presenting GIS.* John Wiley & Sons, West Sussex, England, 149-169.

Chen B and Graddy EA, 2010, The effectiveness of nonprofit lead-organization networks for social service delivery. *Nonprofit Management and Leadership*, 20(4): 405-422.

Hipp JR, Faris RW and Boessen A, 2012, Measuring 'neighborhood': Constructing network neighborhoods. *Social Networks*, 34: 128-140.

Putnam RD, 2001, *Bowling alone: The collapse and revival of American community*. Simon and Schuster, New York, NY, USA.

Kropczynski J and Nah S, 2010, Virtually networked housing movement: Hyperlink network structure of housing social movement organizations. *New Media & Society*, 13(5): 689–703.

Strauss JR, 2010, Capitalising on the value in relationships: A social capital-based model for non-profit public relations. *Prism*, 7(2): 30-35.

# A GIS-based Approach to Determining Possible Influences on a High Quality of Urban Life

Helena Merschdorf[1], Thomas Blaschke[1], Alexander Keul[2]

[1]Department of Geoinformatics (Z_GIS), Schillerstraße 30, 5020 Salzburg, Austria;
merschdorfhe@stud.sbg.ac.at; thomas.blaschke@sbg.ac.at

[2]Department of Psychology, Hellbrunnerstraße 34, 5020 Salzburg, Austria;
alexander.keul@sbg.ac.at

## 1. Introduction

Determining the urban quality of life (QoL), as well as its inherent effect on human behaviour is becoming an increasingly addressed topic within the social sciences. Many quality of urban life studies have been conducted all around the globe, whereby often a combination of objective and subjective analysis methods is used (Marans & Stimson 2011). Objective indicators express hard facts and figures, generally based on aggregated statistical data; subjective indicators express individual human perception. Both dimensions are valuable in reflecting the QoL of a given town, city, or suburb, and therefore both dimensions are commonly analysed in QoL studies. However, due to their inherently different nature, these dimensions are generally analysed separately and for differing underlying reasons, depending on the nature of the respective QoL study. Yet, in order to comprehensively understand all aspects and determinants of a high, respectively low, QoL at a given location, it is essential to consider both objective and subjective factors. Therefore, it is the aim of this study to combine both dimensions into a holistic, integrated research approach, in order to determine interdependencies between both aspects and derive valuable information regarding patterns of QoL. In this sense our main research question is: 'do subjective and objective QoL indicators significantly correspond to one another statistically, and does their spatial contextualization reveal qualitatively identifiable patterns for the case study area of Salzburg, Austria?'

## 2. Methodology

The first stage of the approach involves designing and selecting indicators for both the subjective and objective dimension, which are later to be compared. These chosen indicators are outlined in Table 1. Next, all necessary datasets are accordingly prepared for analysis, whereby the subjective interview data (N=802) are geocoded and the objective data extracted from larger geospatial datasets. Once this stage of data pre-processing has been completed, both the objective and subjective data is spatially analysed for each respective indicator on the common spatial resolution of a 250m grid. Thereby the subjective data is classified according to satisfaction values and aggregated to 250m grid cells, each of which is assigned the modal value of the input data, and the objective data is classified according to its respective density or coverage values.

Table 1: alignment of the chosen indicators and their respective underlying data

| Objective Indicators | Subjective Indicators |
|---|---|
| **Green Coverage** – a 100m raster dataset utilizing a green index from 0 to 1, whereby 0 indicates a concrete jungle and 1 indicates complete green coverage | **Green Satisfaction** – interview item "Do you live in a green neighbourhood"? Grading according to Likert scale: 1 (very) to 4 (not at all) |
| **Housing Density** – a 100m raster dataset containing absolute values regarding the units of housing per cell | **Housing Satisfaction** – interview item "Are you satisfied with your house/ apartment"? Grading according to Likert scale: 1 (very) to 4 (not at all) |
| **Public Facility Distances** – a feature dataset containing distance values for various types of facilities, including schools, kindergartens, parks, sports fields, restaurants, etc., used to create a raster and subsequently extract distance values for each cell to the following facilities: Sports fields, playgrounds, parks, bars and daily shopping facilities | **Public Facility Satisfaction** – composite indicator composed of the following interview items (utilizing the mode values):<br>• "Does your district have sports fields"?<br>• "Does your district have playgrounds"?<br>• "Does your district have parks"?<br>• "Does your district cater for daily shopping needs"?<br>Grading according to Likert scale: 1 (very) to 4 (not at all) |
| **Educational Establishments** – a feature dataset containing distance values for schools and kindergartens, used to create a raster and subsequently extract distance values for each cell to the nearest educational establishment | **Educational Attainment** – interview item "Does your district have good kindergartens and schools"? Grading according to Likert scale: 1 (very) to 4 (not at all) |
| **Public Transport Facilities** – a feature dataset containing distance values for bus and train stations, used to create a raster and subsequently extract distance values for each cell to the nearest public transport node | **Public Transport Satisfaction** – interview item "Does your neighbourhood offer public transport facilities"? Grading according to Likert scale: 1 (very) to 4 (not at all) |

Once all subjective and objective classifications were completed, the correlations between the indicators were analysed in SPSS (v. 16.0) utilizing the nonparametric Kendall's tau-b correlation measure. The measure of correlation serves as an indication as to the statistical relationship between any two variables, whereby one variable is dependent on the other. In this case, the Null Hypotheses (H0) states that there is no linear relationship between the subjective and objective variables. The coefficient values range between -1, indicating a perfect (linear) negative correlation, 0, indicating no correlation, and 1, indicating a perfect (linear) positive correlation.

## 3. Results

The findings of this empirical research suggest that there are statistically significant correlations (Table 2) between several of the objective and subjective indicators, and certain patterns which can be detected through visual interpretation. The correlations are of similar strength to those determined in a previous study utilizing the same subjective data on disaggregated address-based data (Keul et al. 2013), although in this case being aggregated to 250 x 250m cells, allowing for the visual interpretation of patterns. Most of these correlations were found between pairs of subjective or objective indicators, however, some additional correlations between both dimensions were found (Table 2). The highest correlation between objective and subjective data was detected for the indicator 'Green Spaces' (.331), which is also reflected in the patterns depicted in Figure 1. Figure 1 shows that objective green space availability and subjective green space satisfaction are both tendentially higher in the southern districts of Parsch and Aigen, while the opposite is true for the central-northern districts of Lehen, Schallmoos and Elisabeth-Vorstadt.

Table 2: Correlation matrix for all subjective and objective indicators

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Green Spaces (Objective) | Correlation | 1 | | | | | | | | | | | |
| | Sig. (2-tailed) | | | | | | | | | | | | |
| 2. Green Spaces (Subjective) | Correlation | ,331** | 1 | | | | | | | | | | |
| | Sig. (2-tailed) | ,000 | | | | | | | | | | | |
| 3. Housing Density (Objective) | Correlation | ,142** | -,004 | 1 | | | | | | | | | |
| | Sig. (2-tailed) | ,001 | ,936 | | | | | | | | | | |
| 4. Housing Satisfaction (Subjective) | Correlation | ,019 | ,155** | -,104* | 1 | | | | | | | | |
| | Sig. (2-tailed) | ,713 | ,009 | ,045 | | | | | | | | | |
| 5. Public Facilities (Objective) | Correlation | -,177** | -,080 | -,034 | -,043 | 1 | | | | | | | |
| | Sig. (2-tailed) | ,000 | ,120 | ,445 | ,414 | | | | | | | | |
| 6. Public Facilities (Subjective) | Correlation | ,001 | ,185** | -,028 | ,237** | -,078 | 1 | | | | | | |
| | Sig. (2-tailed) | ,986 | ,001 | ,574 | ,000 | ,123 | | | | | | | |
| 7. Educational Establishments (Objective) | Correlation | -,098* | -,054 | -,024 | -,116* | ,165** | -,064 | 1 | | | | | |
| | Sig. (2-tailed) | ,026 | ,291 | ,583 | ,026 | ,000 | ,194 | | | | | | |
| 8. Educational Attainment (Subjective) | Correlation | ,123* | ,214** | -,083 | ,206** | -,104* | ,215** | ,038 | 1 | | | | |
| | Sig. (2-tailed) | ,015 | ,000 | ,103 | ,001 | ,047 | ,000 | ,457 | | | | | |
| 9. Public Transport (Objective) | Correlation | -,301** | -,078 | -,182** | -,080 | ,080 | -,020 | ,125** | -,088 | 1 | | | |
| | Sig. (2-tailed) | ,000 | ,126 | ,000 | ,124 | ,077 | ,683 | ,005 | ,087 | | | | |
| 10. Public Transport (Subjective) | Correlation | -,175** | -,075 | -,113* | ,196** | ,125* | ,211** | ,092 | ,096 | ,232** | 1 | | |
| | Sig. (2-tailed) | ,001 | ,203 | ,027 | ,001 | ,017 | ,000 | ,072 | ,106 | ,000 | | | |
| 11. Elderly Population (Objective) | Correlation | ,252** | ,103* | ,451** | -,016 | -,083 | -,031 | -,129** | -,052 | -,244** | -,192** | 1 | |
| | Sig. (2-tailed) | ,000 | ,042 | ,000 | ,764 | ,064 | ,532 | ,003 | ,309 | ,000 | ,000 | | |
| 12. Elderly Population (Subjective) | Correlation | ,071 | ,112 | -,046 | -,018 | -,121* | ,034 | ,068 | -,013 | -,015 | -,025 | ,001 | 1 |
| | Sig. (2-tailed) | ,158 | ,054 | ,359 | ,769 | ,019 | ,548 | ,181 | ,819 | ,762 | ,672 | ,991 | |
| 13. Overall Subjective QoL | Correlation | ,134** | ,377** | -,115* | ,403** | -,017 | ,219** | -,106* | ,326** | ,042 | ,135* | ,051 | ,097 |
| | Sig. (2-tailed) | ,009 | ,000 | ,026 | ,000 | ,741 | ,000 | ,041 | ,000 | ,422 | ,024 | ,327 | ,104 |

N = 239; **. Correlation is significant at the 0.01 level (2-tailed); *. Correlation is significant at the 0.05 level (2-tailed)

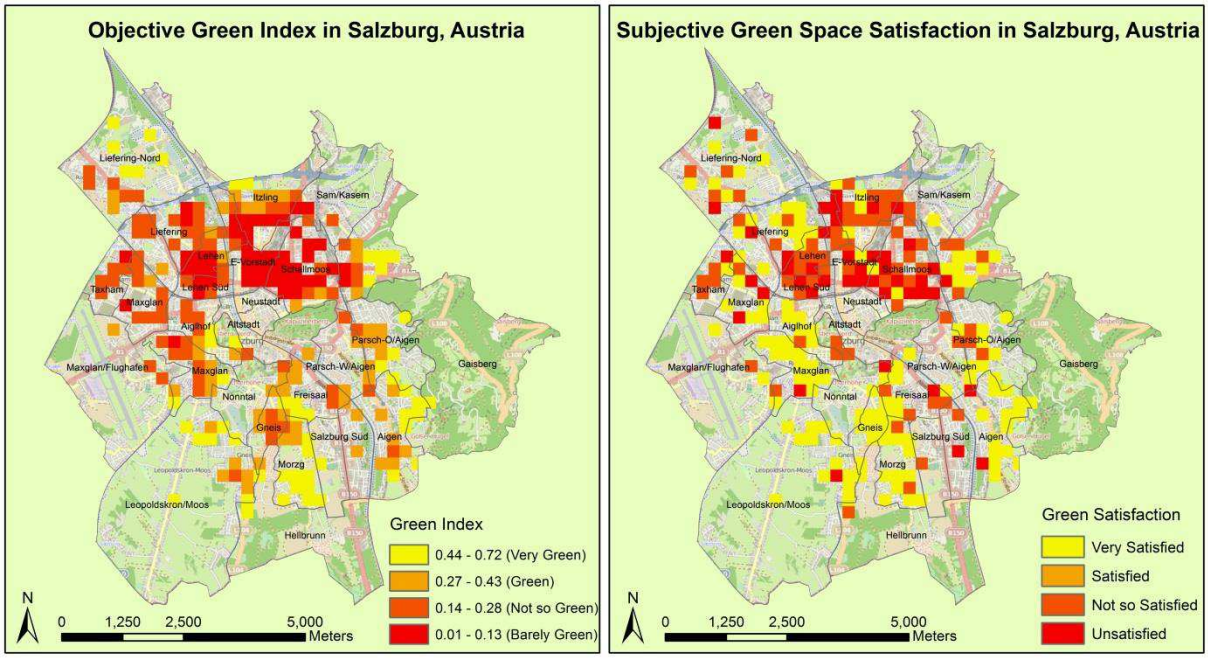



Figure 1: Visual Comparison of Objective and Subjective Green Space Availability

## 4. Conclusion

In alignment with previous attempts to statistically compare objective and subjective indicators of QoL, this study reveals that there are only slight correlations between both

dimensions. This can be attributed to several factors, which basically boil down to the uncertainty inherent in the indicator selection, composition and classification process. For instance, the accuracy of the objective indicator scale range is vastly compromised through the classification into 4 classes, in alignment with the subjective classification, which eliminates the possibility of perceiving fine regional differences. This effect is strengthened by the large cell size which was also chosen due to the availability of the subjective data. If the objective data was classified using a different classification method, the outcome might also vary slightly. Furthermore, the objective indicators do not take into account all factors which may play into the subjective perception, but merely those which are quantifiable and map-able. For instance, the indicator 'public transport' takes into account distance to stations, but not regularity of busses, number of reachable destinations, timeliness, ticket price, and so on. As such, the indicator 'Educational Attainment' doesn't consider quality of education, but rather only distance to the nearest educational facility. These facilities are aggregated in conformity to the subjective data, leaving no distinction between kindergartens, primary schools and secondary schools, which, however, in practice need to be treated separately due to their different requirements, e.g. while a kindergarten needs to be in closer proximity due to the child's inability to get there un-aided, a secondary school can be located further away as long as it is reachable by public transport. Another such example is the indicator of 'Housing Satisfaction', which takes into account the housing density, however, doesn't consider aspects such as housing unit size or interior design components, which are not comprehensively ascertained and available.

These aspects demonstrate that, in order to make it comparable, the objective data is oftentimes strongly compromised in terms of accuracy, which results in a large amount of uncertainty involved in the comparison process. In order to not compromise the objective data, and for it to still be comparable to the subjective data, tens of thousands of interviews would need to be conducted to attain a similar level of spatial coverage, each of which is graded on a scale of at least 0 – 10, rather than 0 – 4, which is logistically next to impossible, and definitely unpractical.

The scientific discussion about the possibility and validity of quantitative–qualitative (nomothetic–ideographic) statistical comparisons in geoinformatics is at least 30 years old. Dale (1980) criticised contradictory quantitative-qualitative results creating a subjective-objective gap, and argued this was the case because of a lack of clear definitions, and because clear relations between indicators and domains were missing. Foo (2000) resumed that subjective indicators had lower measurement reliability, but a higher validity, compared to objective indicators. Cummins (2000) found quantitative-qualitative data to be poorly correlated, but assumed that the social system „homeostatically" maintains subjective QoL within a narrow range.

In this sense, the tendentially lacking statistical coherency between subjective and objective indicators has been outlined in many studies (see related work), yet nevertheless some statistical relationships were successfully identified, despite the spatial aggregation of the data to 250m cells. Considering the incomprehensiveness of the subjective dataset, and the spatial aggregation of both the objective and subjective data, the strength of the detected inter- and crossmodal correlations is notable.

Aside from the correlations detected, the spatial contextualisation of subjective indicators allowed for a graphical depiction, from which patterns could be deduced. This qualitative view is a valuable added insight to the 'hard' statistical (quantitative) analysis, as it allows local knowledge to be integrated and spatial phenomena to be analysed.

Therefore, locating subjective indicators in the same spatial context as the objective indicators gives two advantages: First, it allows a statistical correlation analysis between subjective and objective indicator values to be conducted, which supports the quantitative assessment of dependencies between indicators. Second, it allows a graphical depiction of

both subjective and objective indicators, which supports qualitative deduction of patterns and trends from regional differences that may be neglected in a statistical analysis. In this sense, this study demonstrates the value of locating subjective indicators in a spatial context and integrating quantitative and qualitative analysis to achieve a more holistic understanding of the characteristics influencing urban QoL.

## References:

Cummins, R. A., 2000. Objective and subjective quality of life: An interactive model. Social Indicators Research, 52(1), 55-72.

Dale, B., 1980. Subjective and objective social indicators in studies of regional social well-being. Regional Studies 14(6), 503-515.

Foo, T.S., 2000. Subjective assessment of urban quality of life in Singapore (1997-1998). Habitat International 24, 31-49.

Keul, A.G., Brunner, B. & Spitzer, W. (In press). Wohlbefinden in einer Stadt. Geoinformatik und Prädiktoren subjektiver Lebensqualität in Salzburg. [German, Well-being in a city. Geoinformatics and predictors of subjective quality of life in Salzburg]. Umweltpsychologie.

Marans, R. W. & Stimson, R., 2011. An overview of quality of urban life (pp. 1-29). Springer Netherlands.

# POETS – Python Open Earth Observation Tools

T. Mistelbauer[1], M. Enenkel[1], W. Wagner[1]

[1]Vienna University of Technology, Department of Geodesy and Geoinformation, Gußhausstraße 27-29, 1040 Vienna, Austria
Email: {thomas.mistelbauer; markus.enenkel, wolfgang.wagner}@geo.tuwien.ac.at

## 1. Introduction

While datasets obtained via satellite-based sensors can be powerful, their use in operational applications is still often quite limited. Different data formats, grids, temporal and spatial resolutions, etc. complicate the exploitation of information for many end users. Consequently, the full potential of earth observation is far from being fully exploited. The Python Open Earth Observation Tools (POETS) are a user-friendly open-source toolbox that is capable of accessing, analysing, processing and storing various datasets. It is programmed in Python and allows the integration of individual modules for different users. An example for such a module might be the combination of datasets on rainfall, temperature and soil moisture for a combined drought indicator. Therefore, POETS needs to be capable of handling both historical and near real-time datasets that can be displayed as time series or corresponding images. Instead of visualizing pre-processed illustrations every output is processed on-the-fly. This way disk space can be saved. Users can either access the toolbox from a stand-alone PC or via mobile applications. Pre-configured modules will be available. However, users can also develop their own packages or collaborate with researchers to create more advanced solutions for individual requirements.

## 2. Goals

The major goal of POETS is to simplify the process of collecting, managing and visualizing geospatial data. It is designed to take as much work as possible off the hands of the users, who in the simplest case only need to provide access information to the data sources. Thus, the software prerequisites are narrowed down to only an installation of Python in version 2.7 (van Rossum 1995).

Since there exist several open source libraries for geospatial data handling such as GDAL or pyresample, POETS does not intend to compete against them, but rather to build upon them and use their functionalities as far as possible. Providing these functionalities of data management and visualization as base framework, it is intended to leave the opportunity to further manipulate the data by implementing individual python modules, as discussed in the example in section 3.1.

## 3. Methods

POETS is developed in Python using a NetCDF file (Rew and Davis 1990) for data storage. It uses specific Python packages for managing and manipulating geospatial data such as pyresample and pytesmo (Paulik et al. 2014).

The schema of POETS can be divided into 3 segments, the data providers, the toolbox and the users. The toolbox itself can further be split into distinct python modules where each module is determined to perform a very specific task. Figure 1 illustrates the data flow and processing steps from the data acquisition to the final data visualization on the supported devices.

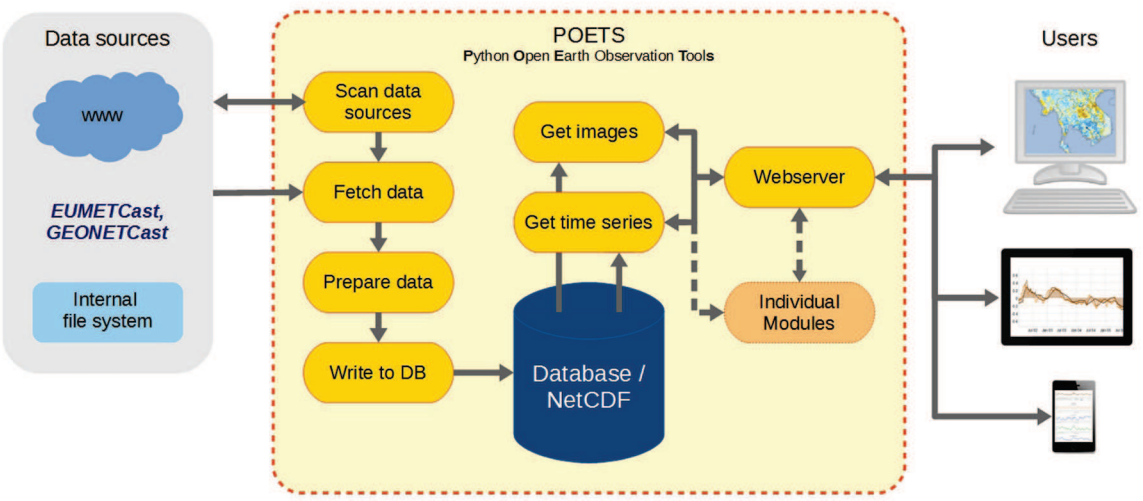



Figure 1 Schema of POETS.

POETS is designed to support various data resources providing data in diverse formats (e.g. NetCDF, HDF5) presuming standing access to online resources (via FTP/SFTP, HTTP) or to internal data repositories. A frequently executed python module monitors the given resources and orders an import module to fetch new data if available. The frequency of that scanning process can be set to any value from one minute up to one month. In a next step, the incoming data are resampled to a predefined grid and temporal resolution using existing python packages. Further, the prepared data are written into the database and thus immediately available to users. The database in this case is a single NetCDF file storing data as images in multiple time layers, which allows sharing the data with other users by simply copying the NetCDF file.

Once the data are imported, third party users can access the data through a web portal, providing both images and time series. The web portal itself is automatically generated by flask, a lightweight web application with a built-in web server (Ronacher 2011). Visualization of the data is realized with the JavaScript libraries dygraphs for time series (http://dygraphs.com, retrieved July 29, 2014) and OpenLayers for images (Jansen and Adams 2010).

Each of the modules shown in Figure 1 can be called separately from any Python code, once the POETS package is installed. Thus, it is possible to only use the download or the resampling routines if nothing else is needed from the package. Figure 2 shows an example of how to import and use modules of POETS in Python separately. In this example, a TIF file is clipped and resampled to the shape of Austria.

```
1 from poets.image.resampling import resample_to_shape
2
3 source_file = 'test_file.tif'
4
5 image, lon, lat = resample_to_shape(source_file, country='AU')
```

Figure 2 Example for the single use of the *resample_to_shape* routine.

### 3.1 Example: Combined Drought Index

The idea for POETS developed during the implementation of a Combined Drought Index in Sout-East Asia (CDI in SEA). The goal of this project was to create a drought index based on various parameters such as precipitation, temperature, vegetation status and soil moisture, as further described by Enenkel (2014).

The tool developed in the CDI in SEA project covered the functionality as shown in Figure 1 lacking the possibility of automated data import and –preparation. The python module responsible for on-the-fly calculation of the Combined Drought Index therefore is an individual module as shown in Figure 1. The CDI in SEA portal can be seen as prototype of POETS and is available at http://geo.tuwien.ac.at/cdi (retrieved July 29, 2014).

## 4. Conclusion and Outlook

POETS is developed as an open-source toolbox for managing and visualizing geospatial data. It is optimized for regional application using a rather coarse spatial and temporal resolution, making it lightweight and resource friendly. However, presuming high performance hardware, it also supports high resolution grids and allows the use of higher temporal resolution.

Once set up, POETS provides data at one specific spatial and temporal resolution. However, it is foreseen to implement both spatial and temporal resampling of the data on-the-fly to provide data at multiple resolutions. It is also planned to provide an extensive application programming interface (API) allowing third party users direct read only access to the database without using the web-portal. Further, it is foreseen to provide the option to use relational database management systems like PostgreSQL for data storage, as well as providing image data via WMS.

Users have the possibility to adapt and develop the tool to their needs by extending the POETS framework or by implementing new Python packages built upon POETS. The source code of this project is hosted at GitHub and can be found at https://github.com/TUW-GEO/poets.

## References

Enenkel M, 2014, Improvement of the Combined Drought Index in South-East Asia, *Final Report for the CDI in SEA Project.*

Jansen M, Adams T, 2010, OpenLayers – Webentwicklung mit dynamischen Karten und Geodaten, *Open Source Press,* ISBN 978-3-937514-92-5.

Paulik C, Steiner C, Hahn S, Melzer T, Gruber A and Wagner W, 2014, Open Source Toolbox and Web Application for Soil Moisture Validation, submitted to *International Geoscience and Remote Sensing Symposium 2014*, Québec, Canada.

Rew R K, Davis G P, 1990, NetCDF: An Interface for Scientific Data Access, *IEEE Computer Graphics and Applications*, Vol. 10, No. 4, pp. 76-82.

Ronacher A, 2011, Opening the Flask: How an April Fools' Joke became a Framework with Good Intentions, http://mitsuhiko.pocoo.org/flask-pycon-2011.pdf, retrieved July 29, 2014

Van Rossum G, 1995, Python Tutorial, *CWI Report CS-R9526.*

# A Comparative Study on the Spatial Statistical Models for the Estimation of Population Distribution

D. R. Oh[1], C. S. Hwang[1]

[1]Department of Geography, Kyung Hee University, 26 Kyungheedae-ro, Seoul 130-701, South Korea
Email: {droh; hcs}@khu.ac.kr

## 1. Introduction

Sometimes we need the spatial data with a finer resolution than the enumeration units such as administrative boundaries for which we have attribute data available. This study aims to estimate the population distribution with more precise spatial resolution using geo-statistical methods. In order to do this, we applied four geo-statistical models and compared the results with each other. Four models are (1) regression model, (2) regression-kriging (RK) model, (3) ordinary kriging (OK) model and (4) co-kriging (CK) model.

Among these models, RK model is generally used to estimate the spatial distribution of soil salinity or soil chemicals (Eldeiry and Garcia 2010), which distribution is closely related to certain spatial factors. We investigate the applicability of RK model to estimate the population of the metropolitan area with the high population density because RK model is suitable to predict the subtle changes. In this respect, we considered the properties and the relations between the size of residential area and the population size for the study area (Dongdaemun-gu and Jungnang-gu) with high population density in Seoul.

The theoretical foundation for the study is a dasymetric interpolation mapping (Eicher and Brewer 2001, Mennis and Hultgren 2006). We used the land use data from the national mapping agency as the ancillary data (the target data in the dasymetric mapping). And the smallest administrative areal unit from the census data was used for the source data. Statistical analyses for the study were performed in the R 2.15.2 and then the ArcGIS 10 was used for the spatial analysis and mapping.

## 2. Approach

### 2.1 Dasymetric Interpolation Mapping with RK Model

The RK model involves the linear regression model that analysis the whole drift/trend between the variables and the kriging model that interpolates the residuals by regression (Hengl et al. 2003). For example, the RK model can be written as equation (1).

$$\widehat{z_{RK}}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) \tag{1}$$

Suppose that $\widehat{z_{RK}}(s_0)$ is the predicted population density at location, $s_0$, $\hat{m}(s_0)$ is the predicted value from the regression model, and $\hat{e}(s_0)$ is the predicted value using the kriging model. This equation can also be expressed as equation (2).

$$\widehat{z_{RK}}(s_0) = \sum_{k=0}^{p} \widehat{\beta_k} \cdot q_k(s_0) + \sum_{i=1}^{n} \omega_i\,(s_0) \cdot e(s_i) \tag{2}$$

Where $\widehat{\beta_k}$ is the estimated regression coefficient, $q_k$ is the independent variable, $\omega_i$ is kriging weight and $e(s_i)$ is the residual by the regression model at the location, $s_i$.

The RK model goes through the following process: Step 1. Correlation coefficients were used in order to determine the relationship of the combination of land uses and population density so the combination with the highest correlation coefficients is selected. Step 2. Ordinary least squares (OLS) multiple regression was used to generate the surface of the population density. In performing a regression analysis, if a spatial autocorrelation is shown in the residuals, a generalized least squares (GLS) regression analysis can supplement OLS multiple regression. Step 3. The values from the regression analysis and the kriged residuals are added, the result will be the estimated population distribution.

### 2.2 Dasymetric Interpolation Mapping with Kriging Model

In order to perform the OK and CK model, the covariance matrix and the variogram are calculated with a primary variable (census data) for the OK model, and a primary variable and a secondary variable (the information of residential area) for the CK model. The primary variable and the secondary variable used in the CK model are interrelated spatially. Among the theoretical semi-variogram, the spherical model shows the best fit using the factors of the semi-variogram in every case in this study. Population weight used on the OK and CK model can be extracted from the semi-variogram model. The arithmetic mean of all the points in the grid cell of the estimated population density was calculated for the predictions.

## 3. Results

### 3.1 Results of the Dasymetric Interpolation Mapping

The results of population distribution using models are counted by the census unit for comparison each other. All models have a similar average value of the estimated population compared to the ordinary/original data. However, the maximum value of the population decreased while its minimum value increased in all models. The degree of change is bigger in the kriging model than in the RK model because the kriging model uses distance functions when it estimates new variables. Figure 1. Shows the predicted population distribution in the study area.

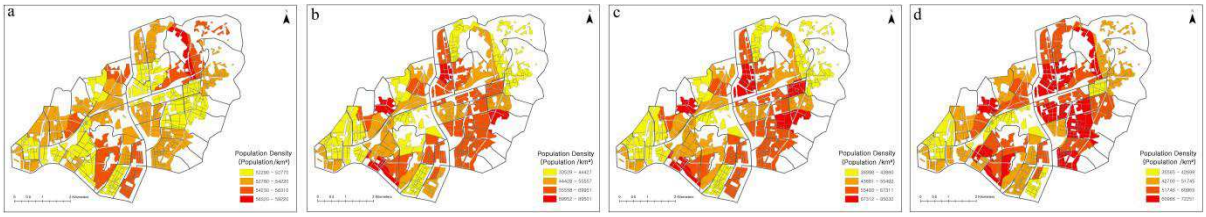



Figure 1: Predicted population density for the study area from
a. regression model, b. OK model, c. CK model, and d. RK model.

### 3.2 Model Evaluation and Validation

Evaluation of the accuracy and validation were the basis on the root mean square error (RMSE), mean absolute error (MAE), goodness of prediction statistic (G statistic) (Kravchenko and Bullock 1999, Guisan and Zimmermann 2000, Eldeiry and Garcia 2010, Kim et al. 2010) and correlation coefficient (ρ). The RMSE can be defined as equation (3) and the MAE can be calculated as equation (4).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{p_i} - p_i)^2}{N}} \tag{3}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}|p_i - \widehat{p_i}|}{N} \tag{4}$$

Where $\widehat{p_i}$ is the predicted value of population at location i, $p_i$ is the ordinary/original value of the population at location i, where i=1, 2, 3, … … . , N.

The effectiveness of the models was measured by the G statistic that can be written as equation (5).

$$\mathrm{G} = [1 - \{\sum_{i=1}^{N}(p_i - \widehat{p_i})^2 / \sum_{i=1}^{N}(p_i - \bar{p})^2\}] \tag{5}$$

Where $p_i$ is the ordinary/original value of the population at location i, $\widehat{p_i}$ is the predicted value of population at location i, and $\bar{p}$ is the mean of the population in the sample area. The model is more efficient when the G statistic has a positive value close to 1. The model is not very efficient when the G statistic has a negative value.

Correlation coefficient (ρ) measures the pattern between the ordinary/original value and the predicted value. Every statistical value of the RK model, the Kriging model, and the regression model, following this order, shows better results. As for the OK model and CK model, both show similar results in this study. Table 1 show the evaluation and validation values on the models.

Table 1. Evaluation and validation values using the regression, RK, OK, and CK models

| | regression | RK | OK | CK |
|---|---|---|---|---|
| RMSE | 4159.51 | 1465.19 | 3851.78 | 3866.07 |
| MSE | 3370.05 | 1001.21 | 3098.42 | 3096.70 |
| G statistic | 0.54 | 0.94 | 0.60 | 0.60 |
| ρ | 0.74 | 0.97 | 0.84 | 0.86 |

### 3.3 Zonal Errors in Population Estimates

In order to investigate the prediction error of a specific unit area, the normalized root mean square error (NRMSE) is utilized. As a result, the RK model's estimation error is large in a low population density zone and small in a high population density zone. Among them, the area with the highest estimation error is revealed to be in the bordering sample areas, even in low population density areas.

Results of estimating the population distribution using the OK and CK model produce high error in boundaries or urban areas with small residential districts in the sample region. Also, the index of local Moran's I shows that a high positive z score for the zone has large errors. Consequently, a very distinct area gap compared to surrounding areas created big errors.

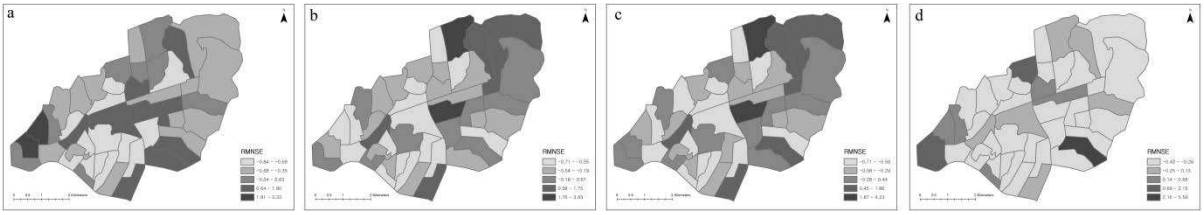



Figure 2: NRMSE values for the study area from a. regression model, b. OK model, c. CK model, and d. RK model.

## 4. Discussion and Conclusions

This study shows that RK model can be an alternative method of estimating a population distribution although the kriging model is frequently used for interpolation. The RK model is suitable for areas with a high population density and a positively high correlation between target data and source data. Therefore, the RK model will be useful for metropolitan areas with a high population density

It is hard to estimate the population distribution using previous weighting methods for dasymetric interpolation mapping, such as the areal weighting method (Goodchild et al. 1993) or the population proportion method (Eicher and Brewer 2001), of the area with land use pattern of the high complexity. The RK model has higher accuracy using the study area compared to the regression, OK, and CK models because the RK model has both advantages of the regression model and the kriging model. And estimated population from the RK method has similar values of descriptive statistics as the ordinary/original data. However, the forms of the model in conjunction with different spatial statistical models involve complicated calculations.

## References

Bracken I and Martin D, 1989, The generation of spatial population distributions from census centroid data, *Environment & Planning A*, 21(4):537-543.

Cromley R G, Hanink M, and Bentley G C, 2012, A quantile regression approach to areal interpolation, *Annals of the Association of American Geographers*, 102(4):763-777.

Eicher C L and Brewer C A, 2001, Dasymetric mapping and areal interpolation implementation and evaluation, *Cartography and Geographic Information Science*, 28(2):125-138.

Eldeiry A A and Garcia L A, 2010, Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using Landsat images, *Journal of irrigation and drainage engineering*, 136(6):355-364.

Fisher P F and Langford M, 1996, Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping, *The Professional Geographer*, 48(3):299-309.

Flowerdew R and Green M, 1992, Developments in areal interpolation methods and GIS, *The Annals of Regional Science*, 26(1):67-78.

Goodchild M F, Anselin L, and Deichmann U, 1993, A framework for the areal interpolation of socioeconomic data, *Environment and Planning A*, 25(3):383-397.

Guisan A and Zimmermann N E, 2000, Predictive habitat distribution models in ecology, *Ecological modelling*, 135(2):147-186.

Hengl T, Heuvelink G, and Stein A, 2003, Comparison of kriging with external drift and regression-kriging, *Technical note, ITC*.

Hengl T, Heuvelink G, and Stein A, 2004, A generic framework for spatial prediction of soil variables based on regression-kriging, *Geoderma*, 120(1):75-93.

Hengl T, Heuvelink G, and Rossiter D G, 2007, About regression-kriging from equations to case studies, *Computers & Geosciences*, 33(10):1301-1315.

Holt J B, Lo C P, and Hodler T W, 2004, Dasymetric estimation of population density and areal interpolation of census data, *Cartography and Geographic Information Science*, 31(2):103-121.

Kim B, Ku C, and Choi J, 2010, Population distribution estimation using regression-kriging model, *Journal of the Korean Geographical Society*, 45(6):806-819.

Knotters M, Brus D J, and Oude Voshaar J H, 1995, A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations, *Geoderma*, 67(3):227-246.

Kravchenko A and Bullock D G, 1999, A comparative study of interpolation methods for mapping soil properties, *Agronomy journal*, 91(3):393-400.

Lee S and Kim K, 2007, Representing the population density distribution of Seoul using dasymetric mapping techniques in a GIS environment, *Journal of the Korean Cartographic Association*, 7(2):53-67.

Liu X H, Kyriakidis P C, and Goodchild M F, 2008, Population-density estimation using regression and area-to-point residual kriging, *International Journal of Geographical Information Science*, 22(4):431-447.

Mennis J, 2003, Generating surface models of population using dasymetric mapping, *The Professional Geographer*, 55(1):31-42.

Mennis J and Hultgren T, 2006, Intelligent dasymetric mapping and its application to areal interpolation, *Cartography and Geographic Information Science*, 33(3):179-194.

Tapp A F, 2010, Areal interpolation and dasymetric mapping methods using local ancillary data sources, *Cartography and Geographic Information Science*, 37(3):215-228.

Wu C and Murray A T, 2005, A cokriging method for estimating population density in urban areas, Computers, *Environment and Urban Systems*, 29(5):558-579.

# On the Assessment of Online Geolocated Social Content for the Identification of Landmarks in Urban Area

T. Quesnot and S. Roche

Center for Research in Geomatics, Université Laval, 1055 Avenue du Séminaire, Pavillon Louis-Jacques Casault, Local 2306
Québec (QC), Canada, G1V 0A6
Emails: teriitutea.quesnot.1@ulaval.ca; stephane.roche@scg.ulaval.ca

## 1. Introduction and Background

Most of the route services provide wayfinding instructions exclusively based on street names. However, it has been demonstrated that for the achievement of a given route, such instructions implied significantly longer delays compared to the landmark-based navigation assistance (Tom and Denis 2003).

Therefore, we know for more than a decade that a route instruction is cognitively suitable when it contains a minimum set of landmarks. More specifically, people's discourse essentially refers to landmarks at choice point areas; i.e. where the traveller has to make a choice (e.g. "turn left"). Two other areas are also concerned: (1) the on-route point portions (landmarks located along the path enable the traveller to ensure he follows the correct route) and (2), the end point, where the presence of landmarks confirms that the traveller has reached the destination (Daniel and Denis 2004).

Based on those studies, researchers have developed systems that automatically detect landmarks. The set of proposed solutions for designing Automatic Landmark Detection Systems (ALDSs) follows the model formalized by Raubal and Winter (2002). This model is applied to the facades of buildings. According to it, the landmarkness is evaluated on the basis of three types of attraction: (1) the visual attraction (e.g. the size of the facade), (2) the structural attraction that mainly refers to the Lynch's structural elements; namely the nodes, boundaries and districts; and (3) the semantic attraction which essentially refers to the cultural and historical significance of the place.

Among all the solutions that have been proposed for designing ALDSs, three have had a significant impact. Firstly, Elias (2003) proposed to identify landmark candidates by using the ID3 algorithm on the building's attributes of a cadastral database. Secondly, Tezuka and Tanaka (2005) have developed a method to mine the web in order to extract landmarks. The task was essentially to evaluate the spatial context of web documents. Unlike Elias's approach, they tried to evaluate the way places are practiced rather than observed. Finally, Duckham et al. (2010) proposed a method that focuses on the top-level categories to which buildings belong instead of their individual characteristics. Thus, their objective was to estimate the landmarkness of a building and not to measure it precisely. This approach is currently the most promising since their algorithm has been implemented on the route service *whereis.com*.

Nevertheless, as explained by Sadeghian and Kantardzic (2008), all of the approaches proposed for designing ALDSs have failed to take into account dynamic parameters, including the measure of objects semantic salience that has often been reduced to the historical and cultural importance of a place; occulting systematically its social dimension yet intrinsic. Richter (2013) argues that this gap can be filled through the use of data generated by the social network users. Few approaches using crowdsourced data have been developed but for the moment, no research has been dedicated to the exploitation of geosocial datasets for the identification of landmarks.

Since the advent of the mobile Internet combined with the development of smartphones, the production of geolocated content from the social web platforms is now commonplace. We are witnessing a real proliferation of spatial data at a massive scale and it does not deal anymore with Goodchild's volunteered geographic information (Goodchild, 2007), but rather with the social location sharing (SLS) phenomenon that distinguish itself by its non-contributory characteristic. This information enable us to access to local geographical knowledge that was previously hardly accessible and measurable. Thus, we argue it should be exploited to enrich ALDSs databases. This is why we propose to evaluate here the potential of geosocial data for the automatic detection of landmarks.

## 2. Method

We propose to assess the potential of SLS datasets through an *in situ* experimentation where two groups of participants will follow a predetermined route in Quebec City (cf. Figure 1). During the travel, they will be asked to choose places they consider to be potential landmarks. The selection will focus on four specific areas: the starting and end points, the choice-point areas and finally the on-route point portions. The first group will leave the site of National Assembly in direction of City Hall while the second will follow the reverse route. Each group will consist of two sub-groups whose members will be designated in function of their degree of familiarity with the study area. Ideally, we would like to have a sex ratio of 1:1 as the wayfinding skills vary from one gender to the other.

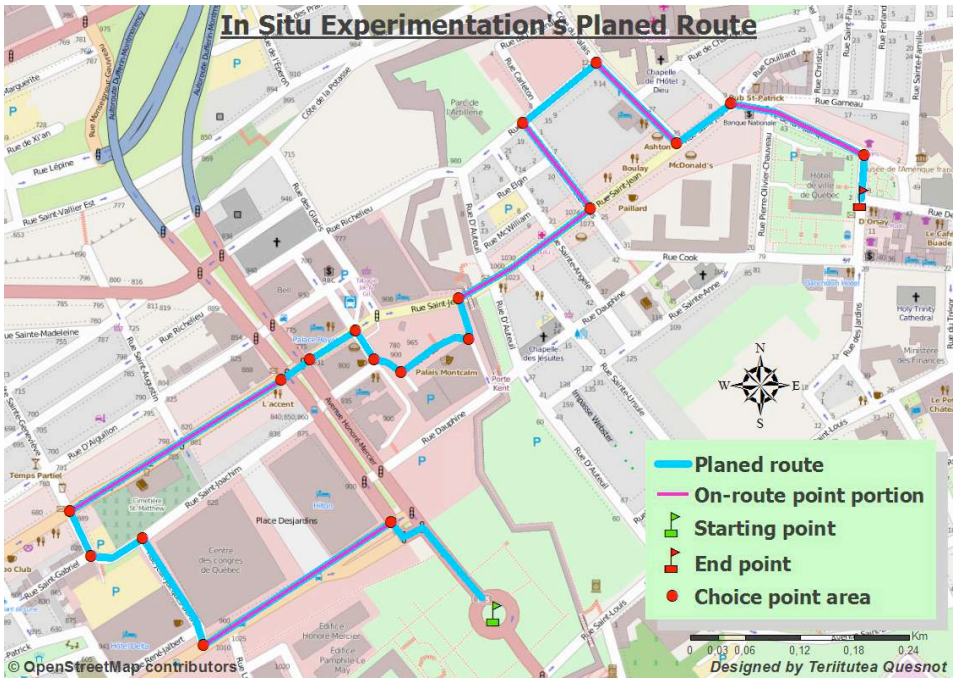



Figure 1. The *in situ experimentation's planed route*.

We consider that the potential of geosocial data in terms of detection of landmarks will be confirmed if a correlation larger than random is found between participant's selection of landmarks and landmarks identified through our approach. We propose to measure the *landmarkness* of a place on the basis of three scores using geolocated data taken from both Foursquare and Facebook's APIs.

The first score corresponds to the *estimated visual salience* (*EVS*). Since the access of precise building's height data is quite difficult, we plan to estimate the visual salience of each Foursquare and Facebook's category by using the criteria established by Duckham et al.

(2010); namely the *place's physical size*, its *prominence*, its *difference from surrounding*, its *daytime and nighttime salience* and finally *its proximity to road*.

The second score is dedicated to the *place's uniqueness* since it is one of the valuable criteria in the detection of landmarks. Specifically, it represents, for a given place *p* located at a checkpoint *ckp*, the ratio between the sum of the places belonging to the category of *p*, and the number of places located at *ckp*; all categories of places combined.

$$\forall n,m>1 : UNQ(p) = \frac{\sum_{i=1}^{n} p_i \in C}{\sum_{j=1}^{m} p_j} \quad (1)$$

where *UNQ* = uniqueness score
*p* = place
*C* = category

The final score represents the *geosocial activity* of a place. This score, applied to Foursquare's check-ins and related information such as tips and "likes", is calculated through the Equation 2. Since the Facebook's API does not provide the distinct number of users who have checked-in at a given place, we propose to calculate the sum of check-ins, "likes" and "talking about" regarding Facebook's geosocial activity score.

$$\forall m \geq 1 : GSA(p)_{4sq} = \frac{CK(p) + LK(p) + TP(p)}{\sum_{j=1}^{m} USR_j(p)} \quad (2)$$

where *GSA* = geosocial activity score
*4sq* = foursquare
*p* = place
*CK/LK/TP* = number of check-ins/likes/tips
*USR* = number of users

At this stage of the project, we propose two assessment options:

- Either compute the *landmarkness score* of a place by calculating the arithmetic sum of these three scores. Therefore, the places selected through our approach will be those with the highest landmarkness score.
- Or rate the quality of theses scores separately by verifying if there is a correlation for each of them with participants' selection of landmarks.

## 3. Expected Results and Impacts

Since it is an ongoing research, we have not yet realized the experimentation. However, two main challenges underlie this research. In fact, the experiment will enable us to check if social networks data are reliable to fill the gaps noted by Sadeghian and Kantardzic (2008) regarding the semantic salience. In this case, designing a geosocial-based ALDS that relies on a user-generated place database would be conceivable. Therefore, data produced by users of social platforms will be concretely exploited through the feeding of a landmark-based

navigation system and not only used for the description of phenomena such as social ties or urban dynamics; as was the case so far. Consequently, we believe the results of this experiment will undoubtedly contribute to the advancement of knowledge in the area of geographic information science and particularly in the field of spatial cognition engineering.

## References

Daniel M-P and Denis M, 2004. The production of route directions: Investigating conditions that favour conciseness in spatial discourse. *Applied cognitive psychology*, 18:57-75.

Duckham M, Winter S, and Robinson M, 2010. Including landmarks in routing instructions. *Journal of Location Based Services,* 4(1):28-52.

Elias B, 2003. Extracting landmarks with Data Mining Methods. In *Proceedings of COSIT 2003*, Kartause Ittingen, Switzerland, 398-412.

Goodchild M F, 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal,* 69(4):211-221.

Raubal M and Winter S, 2002. Enriching wayfinding instructions with local landmarks. In M. J. Egenhofer & D. M. Mark (Eds.), *Geographic Information Science. LNCS 2478*, Berlin, Springer, 243-259.

Richter K-F, 2013. Prospects and Challenges of Landmarks in Navigation Services. In M. Raubal, D. M. Mark & A. U. Frank (Eds.), *Cognitive and Linguistic Aspects of Geographic Space*, Heidelberg, Springer, 83-97.

Sadeghian P and Kantardzic M, 2008. The New Generation of Automatic Landmark Detection Systems: Challenges and Guidelines. *Spatial Cognition & Computation,* 8:252-287.

Tezuka T and Tanaka K, 2005. Landmark extraction: A web mining approach. In *Proceedings of COSIT 2005* , Ellicottville, USA, 379-396.

Tom A and Denis M, 2003. Referring to landmark or street information in route directions: What difference does it make? In *Proceedings of COSIT 2003*, Kartause Ittingen, Switzerland, 362-374.

# Dynamic Visualisation of Complex Spatial Infrastructure Networks

C. Robson, N. Harris, S. Barr, P. James

Newcastle University, School of Civil Engineering and Geoscience, Newcastle-upon-Tyne, NE1 7RU

Email: {c.a.robson1;neil.harris1;stuart.barr; philip.james}@newcastle.ac.uk

## 1. Introduction

Understanding the vulnerability of critical spatial infrastructure networks, such as transport, energy and telecommunications, to perturbations is of significant interest given their increasing importance to our quality of life and economic prosperity. In order to adapt existing critical spatial infrastructure networks to a state of long term resilience it is essential that we understand their current vulnerability due to their spatial and topological organisation.

In this regard, significant progress has been made in developing graph-theoretic failure models that have been applied to study the failure dynamics of both simulated and real spatially complex infrastructure networks. At present such models often characterise the failure dynamics of a network in terms of the changes that occur in their topological structure (Bompard et al. 2011) often on the basis of a particular topological graph metric such as degree distribution, average path length, the size of the giant component or derivatives of these (Holme et al. 2002, Boccaletti et al. 2006). While such metrics are informative and have improved our understanding of the vulnerability of real world spatial infrastructure networks, they do not offer any insight into the explicit vulnerability of the network as a function of its spatial structure. Moreover, they do not provide any obvious means by which the temporal dynamics of the spatial-topological vulnerability of an infrastructure network can be analysed and understood.

In order to address these related limitations we have developed an integrated spatial network failure modelling and dynamic spatial visualisation software framework that allows the real-time visual rendering of the dynamics of network response to a failure model to be presented. An overview of the components of this integrated failure modelling and visualisation framework is shown in Figure 1.

## 2. Infrastructure Network Models

Our tool builds a network from either a shapefile or a postgreSQL/postGIS database table using networkx, a python library for complex network analysis. This in turn is used to build node and edge classes that store attributes including the geometry; the networkx instance is used to perform the analytical network analysis required, while the classes are used to render the features in the visualisation engine. During the network creation stage flow attribute values such as time, distance or cost of traversing the edges are generated and encoded. At the same time, junction nodes that form origin-destination pairs are recognised from a pre-compiled list and used to create a suite of flows for which the shortest paths in terms of the flow attribute of interest are derived and recorded along with a start time for the flow. In order to allow the

dynamic temporal analysis of a spatial failure on a network, the flow along an edge and through a node in the network can be calculated. The assignment of flows and their corresponding time-stamp allows at the failure modelling stage the dynamics of how flows are affected spatially over time to be investigated.

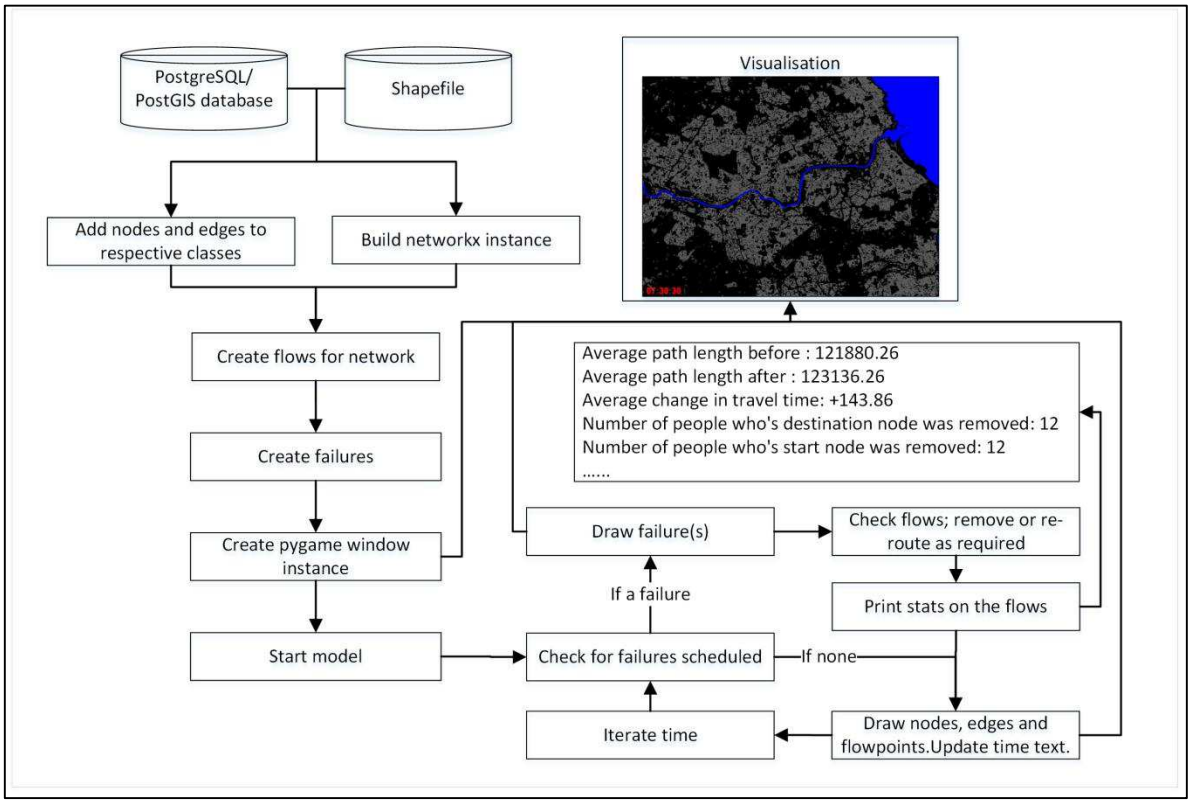



Figure 1: Diagrammatic view of the developed simulation tool summarising the process undertaken from generating the network to identifying a failure and the consequences.

## 3. Failure Modelling

Spatial infrastructure networks are exposed to a wide range of events which can affect component performance, from those related to natural hazards through to failures associated to breakdowns and stress due to demand (Andersson et al. 2005, Demšar et al. 2008). The ability to model such failures has been integrated with the functionality of our tool. To date three forms of failure model have been implemented; (i) spatially random failures, (ii) failures ranked by node degree, and (iii) failures ranked by maximum flow. (ii) allows the node with the highest degree (number of incident edges), thus the most connected, to be removed at any time step, with degree recalculated each time whereas (iii) enables the node with the greatest number of flows over a set time period to be removed, again at any time step. These failure methods can be applied to both nodes and edges (with the exception of (ii) which can only be applied to nodes), with the removal of both being possible within a single simulation; something which has not always been implemented in previous work such as that by Albert et al. (2004) and Crucitti et al. (2004) to give just two examples. When a failure is introduced into the simulation the routes of the generated flows may be affected, thus those which are yet to reach their destination are checked and their routes recalculated if required due to an edge or node on the

route having failed. A summary set of statistics is then produced detailing the effect of the failure on the flows.

## 4. Visualisation engine

The visualisation is based on a pygame (a python library) window instance where both the static background imagery and the network are rendered using their geographic coordinates. At each time step, the size/thickness of the nodes/edges are drawn based on the number of flows which have passed through them in the set time frame, i.e. the last 10 minutes, providing a dynamic temporal view of the network. As a result, where a node/edge has no flows over the time frame it is not rendered on the visualisation. The flows on an edge are visualised at the travel speed of the edge with their position recalculated for each predefined time step of the visualisation. The visualisation is refreshed at every pre-defined time iteration/time step.

## 5. Results

The first example uses the Metro network (an urban light rail system) in the Tyne and Wear region around Newcastle-upon-Tyne, as shown in Figure 2 (below). The first image shows the network in a complete state, whereas the second image shows the state 15 minutes after a node is removed (in red), showing the flow along the affected section of line has severely reduced as no flows can travel through the node, whereas the track on the opposite side of the loop shows a significant increase in flow as would be expected. In this instance, the speed is constant across all edges.

For the road network in Tyne and Wear shown in Figure 3 (below), each road has been assigned a speed appropriate to its road-type. In this example, the node and an edge with the greatest flow over the past 10 minutes at 7:20am and 7:40am are removed respectively, simulating the potential effects of incidents on the major commuting routes. This results in the flows on previously low flow stretches of road increasing dramatically as expected. Although these are only small case studies exemplifying potential applications, larger networks can be analysed to understand their behaviour to perturbations, and the spatial variations in flows throughout the system resulting from multiple simultaneous events on a network.

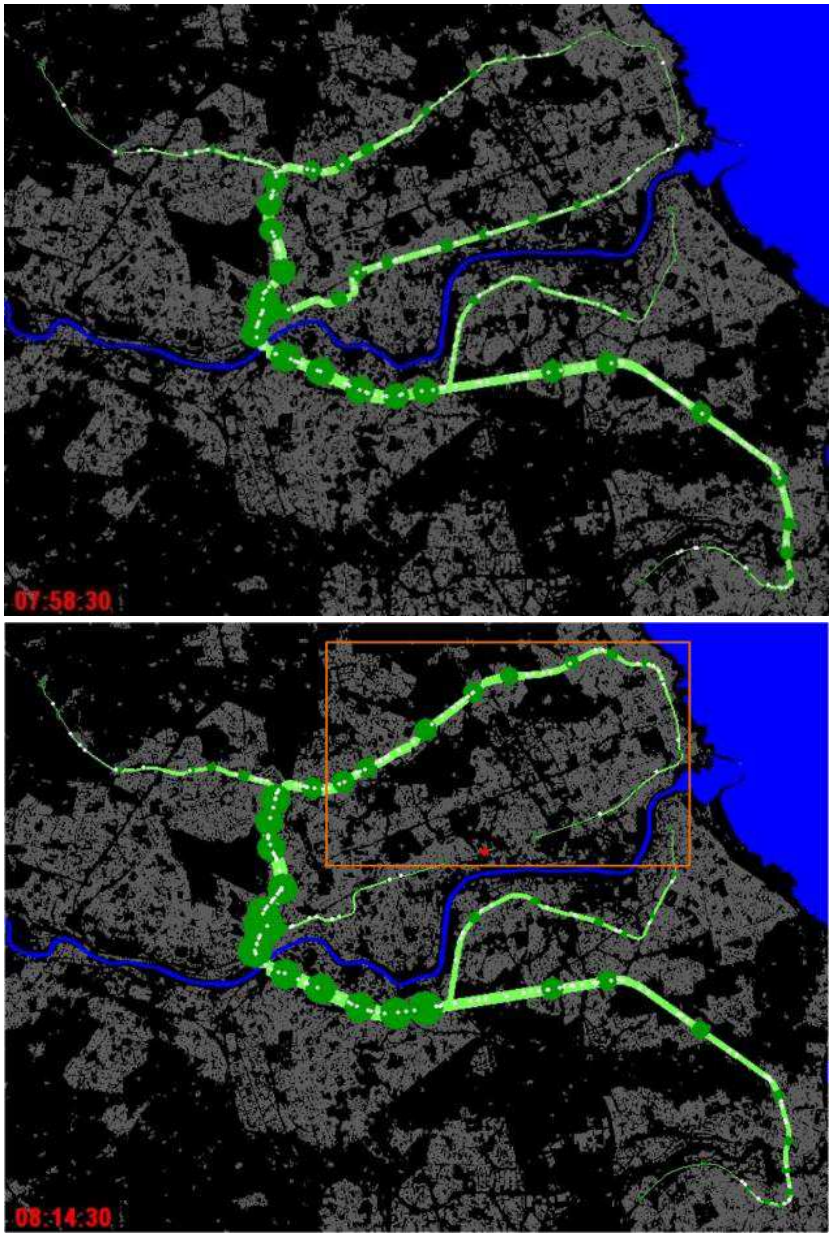



Figure 2: The changing state of the Metro network before and after the removal of a station (marked by a red circle) which also stops flows along the track through the station.

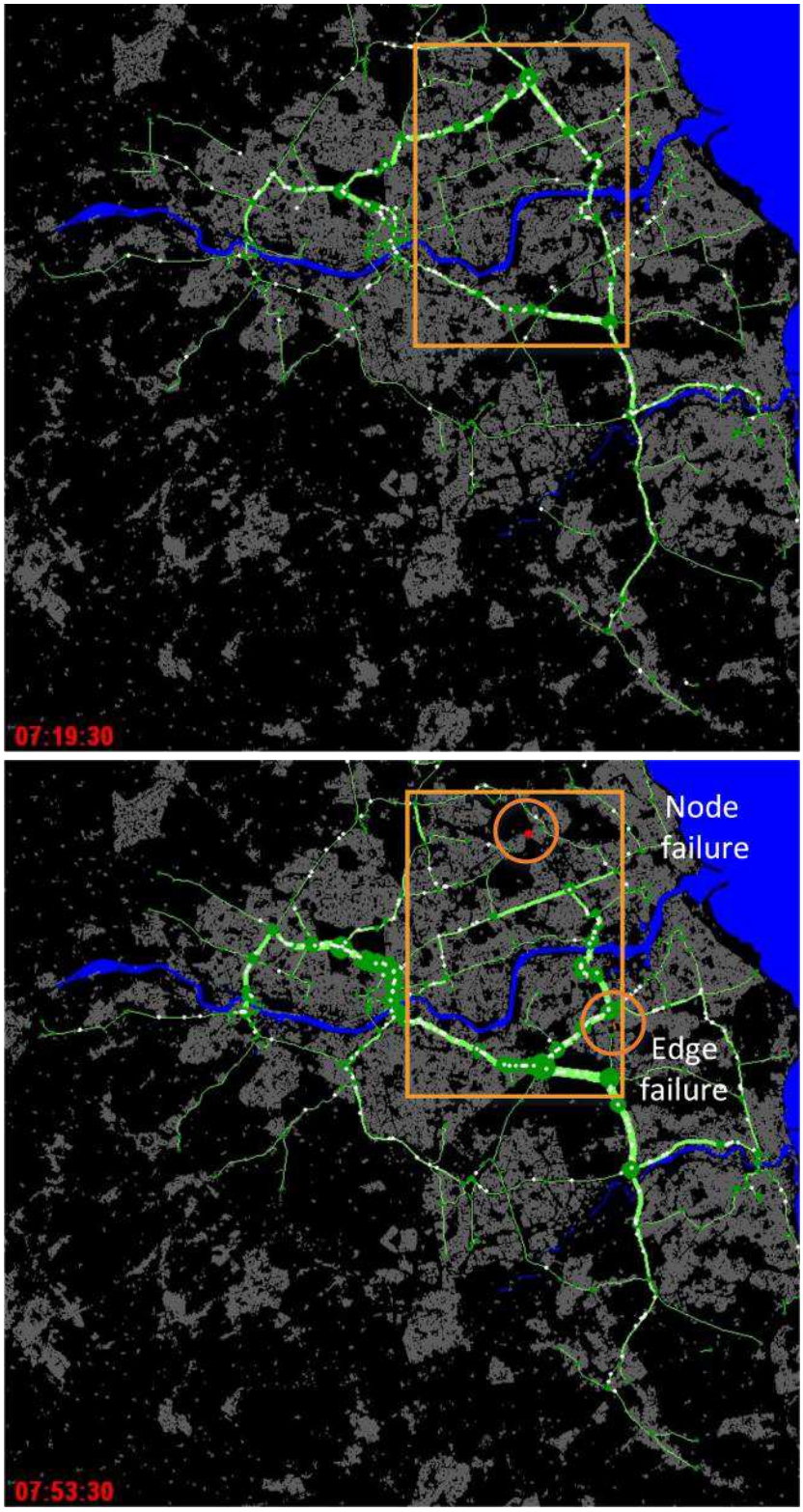



Figure 3: The effect of removing a node and an edge with the greatest number of flows (from the previous 10 minutes) from the road network for the Tyne and Wear region, 20 minutes apart.

## 6. Conclusion

Visualising the spatial impact of perturbations on critical spatial infrastructures has the potential to offer a new insight into the dynamics of the networks which we rely upon for our quality of life and economic prosperity. The tool we have developed allows us to gain a better understanding of network behaviour when exposed to perturbations, enabling changes in flows across an infrastructure to be seen spatially, and allows us to identify those areas where extra demand may be focused, and thus those areas/components which may require adaption to support future events on the network.

## References

Albert, R., Albert, I. and Nakarado, G.L., 2004, 'Structural Vulnerability of the North American Power Grid', The American Physical Society, 69, 1-4.

Andersson, G., Donalek, P., Farmer, R., Hatziargyriou, N., Kamwa, I., Kundur, P., Martins, N., Paserba, J., Pourbeik, P., Sanchez-Gasca, J., Schulz, R., Stankovic, A., Taylor, C. and Vittal, V., 2005, 'Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance', Power Systems, IEEE Transactions on, 20, 1922-1928.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D.U., 2006, 'Complex networks: Structure and dynamics', Physics Reports, 424, 175-308.

Bompard, E., Wu, D. and Xue, F., 2011, 'Structural vulnerability of power systems: A topological approach', Electric Power Systems Research, 81, 1334-1340.

Crucitti, P., Latora, V. and Marchiori, M., 2004, 'A topological analysis of the Italian power grid', Physica A: Statistical Mechanics and its Applications, 388, 92-97.

Demšar, U., Špatenková, O. and Virrantaus, K., 2008, 'Identifying Critical Locations in a Spatial Network with Graph Theory', Transactions in GIS, 12, 61-82.

Holme, P., Kim, B.J., Yoon, C.N. and Han, S.K., 2002, 'Attack vulnerability of complex networks', Physical Review E, 65(5), 1-14.

# Spatial variation of participants' coverage in participatory mapping

Beni Rohrbach[1], Patrick Laube[2]

[1]University of Zurich, Department of Geography, Giscience; Winterthurerstr. 190, 8057 Zürich, Switzerland
Email: benjamin.rohrbach@geo.uzh.ch
[2]Zurich University of Applied Sciences ZHAW, Grüental, 8820 Wädenswil, Switzerland
Email: patrick.laube@zhaw.ch

## 1. Introduction

We present methods to account for the spatial variation of the participants' individual coverage in participatory mapping. Participatory mapping is the process of gathering information in Public Participatory GIS or Participatory GIS. Our studies on participatory mapping revealed a noticeable spatial variance in the participants' coverage when mapping: Some participants even explicitly did not consider some parts of the study area. In this paper we therefore present methods to account for such spatial variation among participants' coverage. The methods were developed and tested around a case study investigating the expected change in extent and of spatial distribution of vineyards in the next decades.

## 2. State of the Art

Sampling strategies in participatory mapping range from casual to purposive. They might be random (Tyrväinen et al. 2007), include volunteers (Brown et al. 2013), represent communities (Alessa et al. 2008), aim at highest diversity of opinions (Debolini et al. 2013) or include selected experts (Yates and Schoeman 2013). However, no known sampling strategy explicitly considers the spatial variation of participants' familiarity within the area of research.

There are several methods to aggregate the spatial expressions of the participants' opinions in participatory mapping. The majority of studies is based on the placement of point markers and subsequent kernel density estimation (Alessa et al. 2008). Other studies ask to map polygons and then count the number of overlaps to highlight converging opinions (Black and Liljeblad 2006), use dot density shading (Montello et al. 2003), or apply a spatial union emphasizing the range of opinions (Morse et al. 2014). But there is no known method to show the range of opinions while still highlighting hotspots.

## 3. Case study: Expected changes in viticulture

### 3.1 Area of investigation

The area of research lies in the alpine canton "Wallis", in southern Switzerland, covering five municipalities (c.f. Figure 1). There, the viticulture is a dominating landscape element, but is expected to change in the near future (Koder 2014). The steep landscape dominated by dry stone walls, as visible in Figure 2, requires much manual labor.

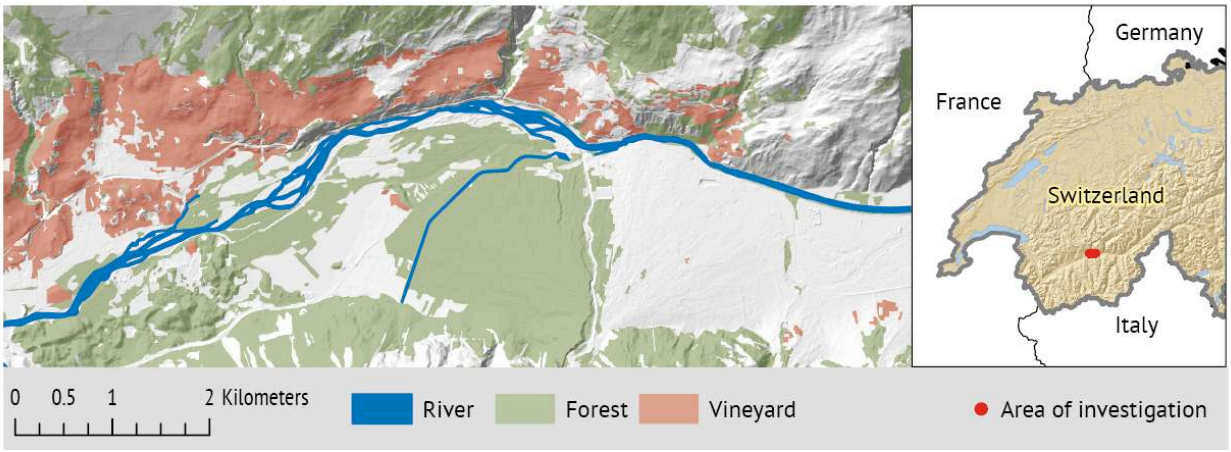



Figure 1: Location of the area of investigation and therein the vineyards

Figure 2: Impression from the area of investigation

## 3.2 Sample

The study investigated the mapping of the expected land-use change in the area. Therefore, we targeted wine-farmers, wine producers, and people that grow grapes as a hobby; approximately 150 candidate participants in the area. A first set of participants was selected in cooperation with a local expert. Additional participants were selected randomly and contacted by telephone, yet others were approached directly in the field. Eventually, 32 participants could be interviewed in person. 5 of the 32 refused to complete the mapping task and one participant covered an area not analyzed further here, resulting in a 26 individual maps containing a total of 288 polygons. Participants were on average 50 years old, with over 25 years of experience in viticulture.

## 3.3 Participatory mapping

The mapping itself was low-tech, low-cost, and reliable, similar to the procedure suggested in Mather et al. (1998). Orthophotos at the scale of 1:5000 on different A3-sheets served as mapping ground, which covered a total area of about 35 $km^2$. To familiarize the participants with the area, they were first asked to mark their own land. Then they were asked to map areas they think will not be used for viticulture anymore in 10 to 15 years from now. The maps were scanned, image processed, georeferenced and then vectorized. The resulting data was processed using FME, qGIS and ArcGIS.

## 4. Aggregation method

First, we assessed the overlap between participants. To do so, the polygons of all participants were overlaid and the number of overlaps was counted. This yields a **density map of opinions** (Figure 3) as known from Black and Liljeblad (2006). Second, we normalized the number of people marking a given area. For normalization, we used the number of participants covering an area in the first place. Therefore, we calculated the area covered by each participant out of all polygons marked by this participant using three different methods: A) The convex hull, B) the concave hull, also known as α-shapes (Edelsbrunner et al. 1983) and C) multiple buffers, resulting in a field-like density surface. In all three methods, the initially mapped shapes were buffered with 50m. This buffer corresponds to the average parcel width in the area and servers as a proxy for the participants' "visual roaming" whilst marking their polygons. The convex hull is a parameter free method. The concave hull requires the setting of an α-value, which we set to 100m after an initial sensitivity study, which roughly corresponds to a "natural" maximal distance between viticulture patches in the area. In the multi-buffer-field method, we used 10 buffers with a distance of 50m each, with the coverage value declining with increasing distance from 1 ("fully covered") to 0.1 ("still somewhat covered"). Summing the coverage of all participants yields the **coverage density map** (Figure 4).

Finally, the density map of opinions was intersected with the coverage density map. Then, we divided the number of opinions by the coverage, resulting in an **agreement map**. Thus, an area covered by 10 and marked 3 participants results 30% agreement among the participants while an area covered by 3 participants and marked by all 3, results 100% agreement. For clarity, areas marked by only one participant are not displayed in Figure 3.

## 5. Results

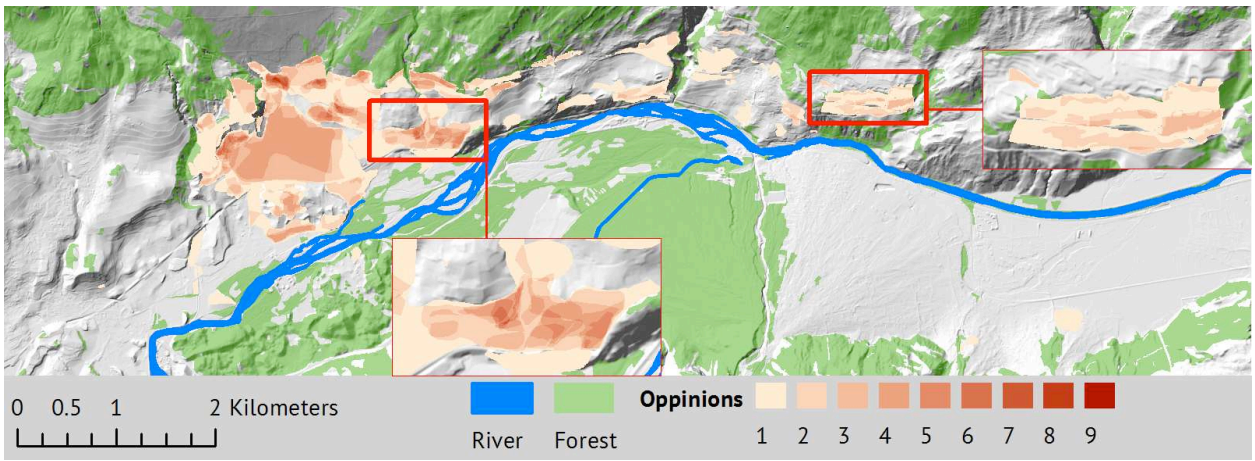



Figure 3: Opinions density map

The data shows a rather low degree of agreement and a strong separation of covered areas among participants. Many participants made statements only about parts of the valley, often about their own municipality. Figure 4 compares the different aggregation methods. The first row illustrates the different methods to estimate the area covered by each participant, and the second row the respective coverage density maps. The third row shows the resulting agreement map. To highlight the differences between the methods, selected parts are zoomed to.

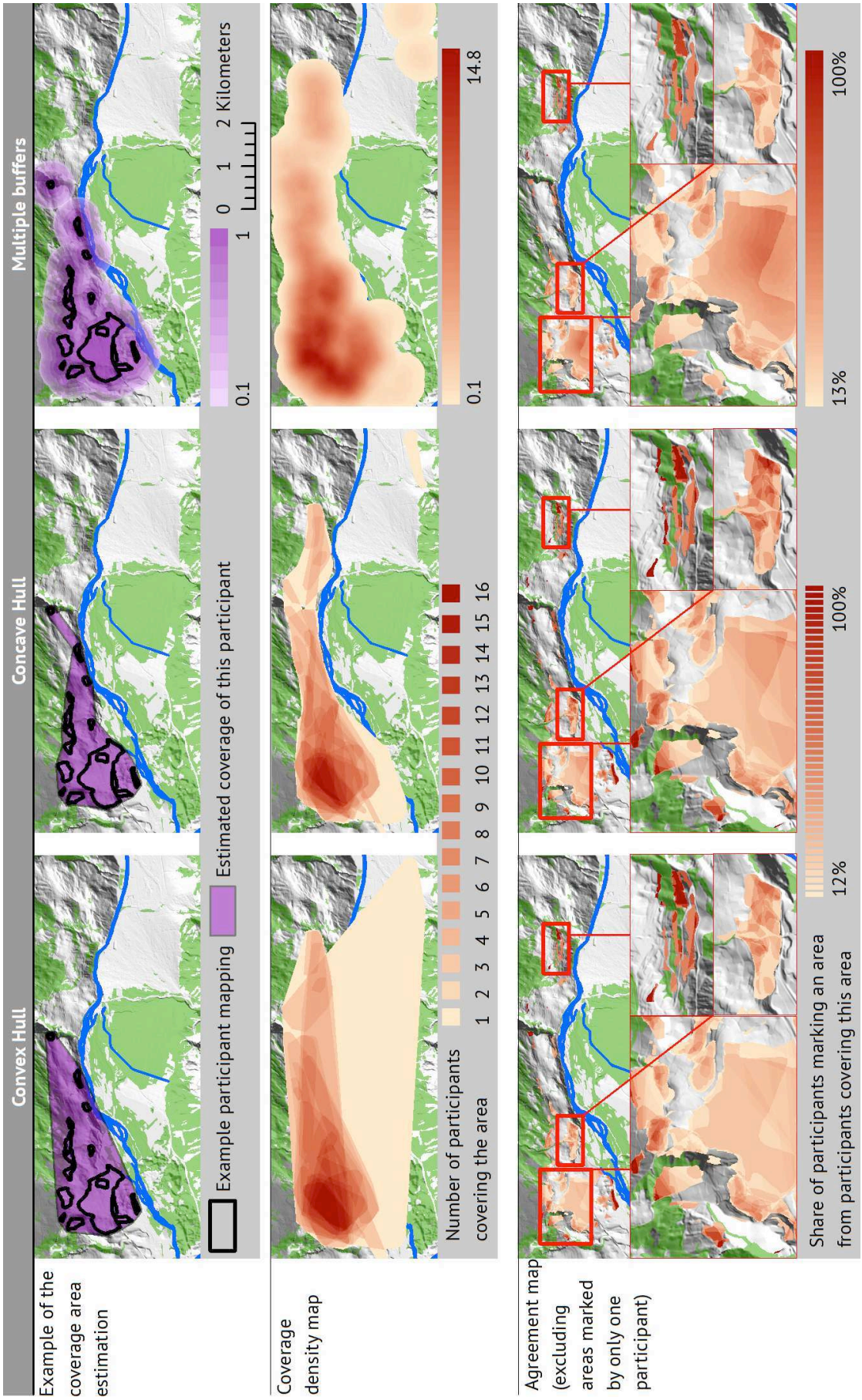



Figure 4: Normalizing the opinion density with the coverage density

## 6. Discussion

Although different in details, the three suggested normalized aggregation methods deliver very similar results (Figure 4). When comparing the normalized aggregation with the conventional opinion density (Figure 3), important differences become evident. Areas that are only mapped by few participants appear to be of little importance in the density map, regardless how many participants covered this area. For example, the areas in the upper right inset map in the third row of Figure 4 indicate high agreement (50-100%) for all normalized aggregation methods, while Figure 3 shows a rather low number of participants marking that area (3-4). Hence, our study helps identifying such potential hotspots that would be overlooked using conventional aggregation.

The proposed methods to calculate the coverage density need further improvement. The convex hull method includes large areas that surely no participant considered. However, as this method is parameter free, it has its advantages. While the multiple buffer approach does smoothen out this effect, this must not be more accurate. Finally, the definition of parameters requires the involvement of domain experts. These parameters may depend on the audience, the communication channel, the mapped objects and the aim of the investigators.

## 7. Conclusions

The contribution of the proposed aggregation method for participatory mapping lies in its ability to show the hotspots of agreement among the participant, while taking the coverage density into account. This work stresses the need to consider spatial differences in coverage, not only for aggregation but as well for sampling.

## Acknowledgments

We would like to thank all the participants and the administration of the protected area “Pfyn-Finges” for their assistance. The University of Zurich funded this work.

## References

Alessa L, Kliskey A, and Brown G, 2008, Social–ecological hotspots mapping: A spatial approach for identifying coupled social–ecological space. *Landscape and Urban Planning*, 85(1): 27–39.

Black A, and Liljeblad A, 2006, *Integrating social values in vegetation models via GIS: The missing link for the Bitterroot National Forest*, Aldo Leopold Wilderness Research Institute.

Brown G, Kelly M, and Whitall D, 2013, Which ‘public’? Sampling effects in public participation GIS (PPGIS) and volunteered geographic information (VGI) systems for public lands management. *Journal of Environmental Planning and Management*, (January): 1–25.

Debolini M et al., 2013, Mapping local spatial knowledge in the assessment of agricultural systems: A case study on the provision of agricultural services. *Applied Geography*, 42: 23–33.

Edelsbrunner H, Kirkpatrick D, and Seidel R, 1983, On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4): 551–559.

Koder W, 2014, Raron kämpft um seine Rebberge. *Walliser Bote*, March 5: 2

Mather R et al., 1998,Aerial Photographs and Photo-Maps for Community Forestry. *Rural Development for Network (RDFN) papers*, Rural Development for Network (RDFN) papers,

Montello D et al., 2003, Where’s Downtown ?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition & Computation*, 3(2-3): 185–204.

Morse WC, Lowery DR, and Steury T, 2014, Exploring Saturation of Themes and Spatial Locations in Qualitative Public Participation Geographic Information Systems Research. *Society & Natural Resources*, (April): 1–15.

Tyrväinen L, Mäkinen K, and Schipperijn J, 2007, Tools for mapping social values of urban woodlands and other green areas. *Landscape and Urban Planning*, 79(1): 5–19.

Yates KL, and Schoeman DS, 2013, Spatial access priority mapping (SAPM) with fishers: a quantitative GIS method for participatory planning. *PloS one*, 8(7): e68424.

# OpenStreetMap data assessment for extraction of urban land cover and geometry parameters required by urban climate modeling

T.E.Samsonov[1], P.I.Konstantinov[2]

[1]Automation Lab, Dept. of Cartography and Geoinformatics,
Faculty of Geography, Lomonosov Moscow State University,
Leninskiye Gory 1, Moscow, Russia, 119234
Email: tsamsonov@geogr.msu.ru

[2]Meteorological Observatory, Dept. of Meteorology and Climatology,
Faculty of Geography, Lomonosov Moscow State University,
Leninskiye Gory 1, Moscow, Russia, 119234
Email: kostadini@mail.ru

## 1. Introduction

More than a half of world's population lives in cities now, and the urban/rural ratio is increasing (World Urbanization… 2014), which deserves increased attention to heavily populated areas. Contemporary meteorological models assimilate information about urban conditions for implementation of scenarios of physical interaction between atmosphere boundary layer and underlying surface (Kusaka et al. 2001). This allows more precise weather and climate predictions (Konstantinov et al. 2014).

During the last two decades significant progress has been made in description of urban environment for urban climate modeling. Required parameters are extracted from satellite imagery and spatial datasets such as city vector geodatabases (Lindberg 2007). However, expensiveness and unavailability of timely data often limits the possibility of urban climate studies. Recently, volunteered geographic information received great attention as a source of information about land cover (Comber et al. 2013). Our research pioneers in the assessment of OpenStreetMap data for possibility of extraction of main parameters of land cover and urban geometry needed for urban climate research.

## 2. Land cover classification

Meteorological (climate) models consider physical characteristics of surfaces to model their interactions with atmosphere. For example, WRF model (Skamarock et al. 2008) uses GLCC 1 km resolution land cover database (Loveland et al., 2000) that contains 24 land cover classes. This information should be refined for urban areas.

Meso-scale models use simple land cover refinements. For example, Kusaka et al. (2001) considers urban and vegetation ratio (implemented in WRF), while Trusilova et al. (2013) differentiates fractional area of urban land and fractional artificial area occupied by buildings. At the same time considering water and green area ratios in experimental high-resolution models facilitates better reproduction of temperature effects above those surface types (Konstantinov et al. 2014). This demonstrates the potential of fine-grained classifications of urban land cover for micro-scale modeling.

Guided by availability of various OSM keys (OpenStreetMap... 2014) we developed reclassification scheme that is close to proposed by Lemonsu et al. (2008) and is presented in Table 1. This classification can then be easily reclassified into more simple parameterizations.

Table 1. Extraction of land cover types from OSM Data

| # | OSM Key | OSM values | Destination class |
|---|---|---|---|
| 1 | building | ALL except bunker / cabin / construction / farm_auxiliary / hut / shed / stable | buildings |
| 2 | waterway | river / riverbank / stream / canal / ditch | water |
| 2 | natural | wetland / water | water |
| 2 | landuse | reservoir | water |
| 3 | natural | tree / tree_row / wood | tall vegetation |
| 3 | landuse | forest | tall vegetation |
| 4 | landuse | orchard / vineyard / scrub / farm / farmland / greenfield | low vegetation |
| 5 | landuse | grass / meadow / pasture | grass |
| 6 | leisure | garden / park | mixed vegetation |
| 7 | surface | ground / earth / dirt / mud / sand | bare ground |
| 7 | natural | bare rock / mud / sand / beach | bare ground |
| 8 | surface | asphalt | asphalt |
| 9 | surface | concrete | concrete |
| 10 | highway | ALL except track / path / footway / bridleway / steps / proposed | roads |
| 11 | landuse | construction / garages, industrial / military / railway | industrial |

In current research we focused on the availability of building data as being the most important, keeping assessment of other land cover types for future investigations.

## 3. Urban canyon geometry

The central concept of urban meteorology is *urban canyon* (Nunez and Oke 1977) which stands for the space between buildings characterized by its width (W), height (H) and length (L) (Figure 1).

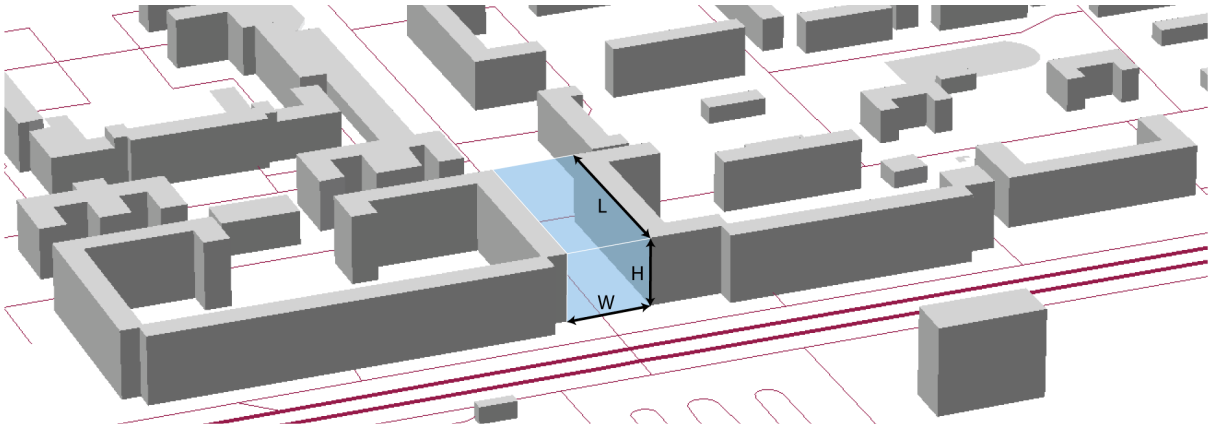



Figure 1. Urban Canyon

Current meso-scale models are not canyon-resolving due to their spatial resolution (~1 km). They assimilate mean parameters of buildings to reconstruct average canyon geometry in every cell. Trusilova's et al. (2013) scheme includes such parameters as building height, height to width ratio of canyons and roughness length for the building–canyon system. Kusaka et al. (2001) uses more sophisticated parameterization that includes street-canyon orientation. Derivation of these parameters requires information about buildings and their heights.

Buildings are coded in OSM data using “building” tag which can be filled by simply “yes” value or the value containing the particular type of the building. There are also two options for coding building heights. The first is “building:height” tag and the second is “building:levels” tag. As detailed information about precise building height is rarely available the second tag is much more common.

We examined the completeness of OSM building data in 29 largest world urban areas that have more than 10 mln inhabitants as of March 2014 (Demographia... 2014). Results are summarized in Table 2. L-ratio reflects how many buildings are attributed with levels data. B-ratio is synthetic index that shows how many buildings are digitized in relation to number of inhabitants and thus reflects the completeness of building geometry.

Table 2. OSM building data availability for world's largest urban areas (> 10 mln people)*

| # | Rank | Country | Name | Population | Area | Density | Buildings | Levels | L-ratio | B-ratio |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **27** | **France** | **Paris** | **10 975** | **2 845** | **3,9** | **2415331** | **2333** | **0,10%** | **220,08** |
| 2 | **8** | **United States** | **New York** | **20 661** | **11 642** | **1,8** | **1103982** | **692** | **0,06%** | **53,43** |
| 3 | **29** | **United Kingdom** | **London** | **10 149** | **1 738** | **5,8** | **532550** | **11240** | **2,11%** | **52,47** |
| 4 | **15** | **Russia** | **Moscow** | **15 885** | **4 662** | **3,4** | **337229** | **58386** | **17,31%** | **21,23** |
| 5 | **28** | **Japan** | **Nagoya** | **10 238** | **382** | **2,7** | **152337** | **768** | **0,50%** | **14,88** |
| 6 | **14** | **Japan** | **Osaka-Kobe-Kyoto** | **17 234** | **3 212** | **5,4** | **249902** | **20692** | **8,28%** | **14,50** |
| 7 | **1** | **Japan** | **Tokyo-Yokohama** | **37 555** | **8 547** | **4,4** | **516715** | **6611** | **1,28%** | **13,76** |
| 8 | **5** | **Philippines** | **Manila** | **22 710** | **158** | **14,4** | **218464** | **699** | **0,32%** | **9,62** |
| 9 | 16 | United States | Los Angeles | 15 250 | 6 299 | 2,4 | 51341 | 416 | 0,81% | 3,37 |
| 10 | 2 | Indonesia | Jakarta | 29 959 | 3 108 | 9,6 | 86109 | 11084 | 12,87% | 2,87 |
| 11 | 20 | Bangladesh | Dhaka | 14 816 | 337 | 44 | 19510 | 14269 | 73,14% | 1,32 |
| 12 | 18 | Thailand | Bangkok | 14 910 | 2 461 | 6,1 | 17120 | 210 | 1,23% | 1,15 |
| 13 | 26 | Brazil | Rio de Janeiro | 11 723 | 202 | 5,8 | 9882 | 791 | 8,00% | 0,84 |
| 14 | 23 | Turkey | Istanbul | 13 187 | 1 347 | 9,8 | 10995 | 54 | 0,49% | 0,83 |
| 15 | 21 | Argentina | Buenos Aires | 13 913 | 2 642 | 5,3 | 10931 | 123 | 1,13% | 0,79 |
| 16 | 11 | China | Beijing | 19 277 | 3 756 | 5,1 | 14099 | 92 | 0,65% | 0,73 |
| 17 | 6 | China | Shanghai | 22 650 | 3 626 | 6,2 | 14749 | 182 | 1,23% | 0,65 |
| 18 | 19 | India | Kolkota | 14 896 | 1 204 | 12,4 | 9674 | 0 | 0,00% | 0,65 |
| 19 | 10 | Brazil | Sao Paulo | 20 273 | 2 849 | 7,1 | 12632 | 2549 | 20,18% | 0,62 |
| 20 | 13 | India | Mumbai | 17 672 | 546 | 32,3 | 10524 | 100 | 0,95% | 0,60 |
| 21 | 17 | Egypt | Cairo | 15 206 | 1 761 | 8,6 | 6069 | 49 | 0,81% | 0,40 |
| 22 | 4 | South Korea | Seoul-Incheon | 22 992 | 2 266 | 10,1 | 8418 | 119 | 1,41% | 0,37 |
| 23 | 9 | Mexico | Mexico City | 20 300 | 2 072 | 9,8 | 3648 | 119 | 3,26% | 0,18 |
| 24 | 3 | India | New-Delhi | 24 134 | 2 072 | 11,6 | 4219 | 41 | 0,97% | 0,17 |
| 25 | 24 | China | Shenzhen | 12 860 | 1 748 | 7,4 | 2073 | 37 | 1,78% | 0,16 |
| 26 | 25 | Nigeria | Lagos | 12 549 | 907 | 13,8 | 795 | 0 | 0,00% | 0,06 |
| 27 | 22 | Iran | Tehran | 13 429 | 136 | 9,9 | 844 | 23 | 2,73% | 0,06 |
| 28 | 7 | Pakistan | Karachi | 21 585 | 945 | 22,8 | 1082 | 17 | 1,57% | 0,05 |
| 29 | 12 | China | Guangzhou-Foshan | 18 316 | 3 432 | 5,3 | 824 | 0 | 0,00% | 0,04 |
| | | | | X10³ people | km² | x10³ people / km² | count | count | Levels / buildings, % | Buildings / 10³ people |

*The list of urban areas and data about population, area and density is taken from (Demographia... 2014).

Results show only 8 cities with high values of B-ratio (bold font) — those having relatively full information about built-up. And only 4 cities (highlighted in green) have

significant value of L-ratio. The most satisfactory results are shown by Moscow city, however even there the completeness of information is not enough for its usage in urban studies, as only 17% of buildings are attributed with levels.

We assessed the quality of OSM building levels in Moscow city using Geocentre Consulting Ltd. database. OSM levels (L) were reduced to heights using H = 4L formula. Heights from both databases was averaged for 262 cells with 200 m resolution. Figure 2 presents scatterplot with reference heights along Y axis and OSM heights along X. Coefficient of determination is $R^2 = 0.77$ for this dependency. This shows satisfactory level of dependency and proves that data can be potentially used in urban climate tasks. However, the similar verification should be done for all other major urban areas in the future.

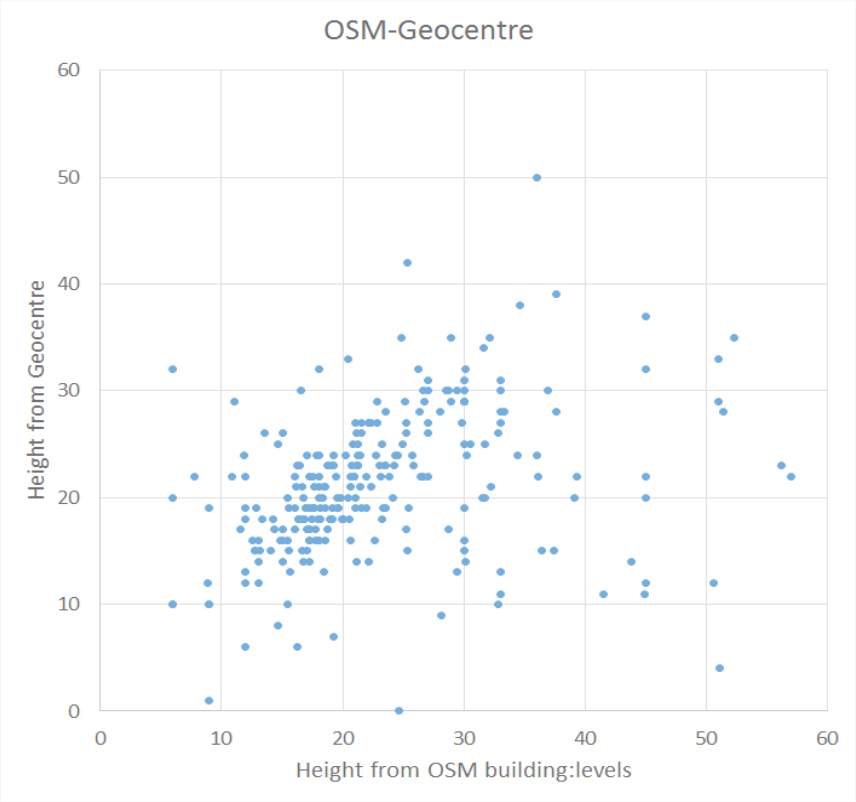



Figure 2. Scatterplot of OSM building heights and referential heights from Geocentre Consulting database (262 cells with 200 m spatial resolution).

## 4. Discussion

In this paper, OSM data is assessed in terms of land cover and urban geometry characterization for urban climate modelling for the first time. A mapping of OSM tags to land classes is proposed. The completeness of OSM building data is estimated over 29 largest world urban areas. Results showed that OSM data is nor ready yet for urban canyon estimations due to incompleteness of buildings and/or their levels attribute. The quality assessment of Moscow OSM building levels show satisfactory correspondence with reference data and thus potential applicability of OSM data in extraction of urban canyon geometry.

## Acknowledgements

This work was supported by RFBR grant 13-05-41306-RGO_a.

# References

Comber AJ, Brunsdon C, See LM, Fritz S, McCallum I, 2013. Comparing Expert and Non-expert Conceptualisations of the Land: An Analysis of Crowdsourced Land Cover Data. LNCS 8116: Proceedings of 11th International Conference COSIT 2013, Springer, p 243-260.

Demographia World Urban Areas (Built-Up Urban Areas or Urban Agglomerations) 10th Annual Edition: March 2014, Wendell Cox Consultancy, Belleville, IL, USA, 127 p.

Konstantinov PI, Varentsov MI, Malinina EP, 2014. Modeling of thermal comfort conditions inside the urban boundary layer during Moscow's 2010 summer heat wave (case-study). Urban Climate (in press). DOI: 10.1016/j.uclim.2014.05.002. 10 p.

Kusaka H., Kondo H, Kikegawa Y, Kimura F, 2001. A simple single-layer urban canopy model for atmospheric models: Comparison with multi-layer and slab models. Boundary-Layer Meteorology, vol. 101, pp. 329–358.

Lemonsu A, Leroux A, and Belair S., 2008. A general methodology of urban land cover type classication for atmospheric modelling. Technical report, Meteo France, Centre National de Recherches Meteorologiques

Lindberg F, 2007. Modelling the urban climate using a local governmental geo-database. Meteorological Applications, 14 (3), 263-273.

Loveland TR, Reed BC, Brown JF, Ohlen DO, Zhu, J, Yang, L., and Merchant, J.W., 2000, Development of a Global Land Cover Characteristics Database and IGBP DISCover from 1-km AVHRR Data:International Journal of Remote Sensing , v. 21, no. 6/7, p. 1303-1330

Nunez M and Oke TR, 1977. The energy balance of an urban canyon. *Journal of Applied Meteorology*, 16, 11-19.

OpenStreetMap Wiki, 2014. http://wiki.openstreetmap.org [accessed 11 Aug 2014]

Skamarock W, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda MG, et al. (2008) A description of the advanced research WRF version 3: NCAR technical note TN–475+STR. National Center for Atmospheric Research Boulder, Colorado, USA, 125 p.

Trusilova K, Früh B, Brienen S, Walter A, Masson V, Pigeon G, Becker P, 2013. Implementation of an Urban Parameterization Scheme into the Regional Climate Model COSMO-CLM. J. Appl. Meteor. Climatol., vol. 52, no. 10, p. 2296–2311.

World Urbanization Prospects, the 2014 Revision: Highlights, 2014. United Nations, Department of Economic and Social Affairs, Population Division: New York, 32 p.

# A Spatial Agent-based Model for Assessment and Prediction of Woodchips Availability for Heating Plants in Austria

Johannes Scholz[1], Peter Mandl[2], Christian Kogler[2], Michael Müller[2]

[1]Research Studios Austria – Studio iSPACE, Schillerstrasse 25, A-5020 Salzburg, Austria
Email: johannes.scholz@researchstudio.at

[2]Department of geography and Regional Studies, University Klagenfurt, Universitaetsstrasse 65-67, 9020 Klagenfurt, Austria
Email: peter.mandl@aau.at, {chris; m9muelle}@edu.uni-klu.ac.at

## 1. Introduction

Energy from renewable sources is a growing trend in Europe and all over the world, as a reduction of greenhouse gases by a certain percentage is a political goal according to the Kyoto Protocol. In order to achieve the goals envisaged in the treaty – a reduction of greenhouse gas emissions by 20% (base year 1990) during 2013-2020 – the propagation of renewable energy sources is one of the strategies followed by politics and administration.

Though in every country the increased usage of renewable energy sources is favoured and sponsored. There is rarely a study that evaluates the effects of the consumption of renewable energy resources with respect to the spatio-temporal dimension on a fine grained level of detail. So far only studies exist that evaluate renewable energy projects on a global/local level - e.g. for a whole country or province (e.g. Arbeitsplattform Wald und Holz in Kärnten 2007, Möller and Nielsen 2007).

The work presented in this paper focuses on effects of wood chip heating plants on the availability of lumber for wood chip production on a fine spatial and temporal granularity. In order to model the "consumption" of timber for heating purposes, an agent-based model coupled with a GIS is employed (Crooks and Heppenstall 2011, Johnston 2013, Van Berkel and Verburg 2012, Mandl 2003). In a generic way the model mimics the competition of several heat plants for biomass. In addition, it includes a basic forest growth model as well as a sequence of forest operations that result in lumber for wood chip production. The scientific question of this paper tries to answer if a spatio-temporal effect of competition for renewable energy resources exists. The paper evaluates the question in the context of the wood chip market which serves as energy source for heating plants in a given test area. The theoretical model is applied to a data set of Carinthia, a province of Austria.

The paper is organized as follows: section 2 elaborates on the test area, the data used in the study and the general approach of the study. The theoretical model, following an agent-based approach, is described in section 3. The first results are given in section 4, followed by a discussion and conclusions.

## 2. Test Area and Approach of the Study

This section elaborates on the study area and the necessary data for conducting some experiments using an agent-based model. The scientific approach of this study is given in a subsequent subsection.

## 2.1 Study area: Province of Carinthia and data used

The test area of this study is the province of Carinthia - the southernmost federal state of Austria. The test area is chosen due to the fact that 57.6 % of the area is covered by forests and several heating plants are existing which compete for the available biomass (i.e. wood).

The data necessary for this study are a road network, spatial raster data indicating forest according to the Austrian Forest Act with 300 m resolution (source: Austrian Ministry for Agriculture & Forestry), forest type map raster data (source: Austrian Ministry for Agriculture & Forestry), growing stock raster data (source: Austrian Ministry for Agriculture & Forestry), a digital elevation model with 25 m resolution and data on of the position and yearly lumber consumption of heating plants in Carinthia (see Figure 1).

The road network and the DEM with 25 m resolution are used to model the distance between heating plants and forest stands – i.e. the biomass to be harvested. The forest data – forest, forest type, growing stock – are necessary to reason about the forest type, tree age, the forest growth and the potential biomass to be extracted from the forest (either as thinning or clear cut, depending on the age of the standing trees). The data on the heating plants – position and yearly biomass consumption are created for the study by the authors.

## 2.2 General Approach of the Study

In order to model the effects of wood chip heating plants on the availability of wood chips based on the available data sources we defined the following spatio-temporal model that is based on an agent-based approach. The model is given in Figure 1 and described hereafter.

In order to model the biomass supply chain a source and a sink has to be defined. The source is the forest, where timber is growing constantly with respect to a forest growth model and the current growing stock. Lumber for wood chip production can only be extracted from the stands based on a given forest operations plan which is based on schedule for thinning operations and clear cuts based on the age of the trees. The sinks are represented by heating plants which generate energy from wood chips.

The reason for modelling the biomass supply with agents traveling in space to collect biomass is explained as follows. The system should model a dynamic “market situation” for a given time period of several decades. This shall give evidence if there is a spatio-temporal effect on the biomass availability – i.e. the market – for heating plants. Hence, each agent is able to interact and alter the environment accordingly – by collecting the timber of thinning and clear cut operations and converting them into biomass for energy purposes. To mimic the behaviour of heating plant managers – whose primary target is to minimize the transport costs of biomass – we assume that each manager collects available timber close to the location of the heating plant with respect to the street network. Due to the fact that there are a number of different heating plants in the province of Carinthia, there is a market situation where heating plants compete for close available biomass, we follow an agent-based approach to mimic the market situation present in the universe of discourse. In this paper we intentionally do not model the real market as such, as there are no studies on the behaviour of the biomass market as such. Thus, we model the market by agents, with a given behaviour, which follow the basic rationale any market participant follows. In addition, the study focuses on modelling agents on a fine grained level of detail, but restricts the analysis to the macro-level which gives a global overview of the outcomes of the biomass market.

Each heating plant gets a “prospector” - i.e. an agent - that searches for available biomass, which is produced in the forest with respect to the forest operations schedules. The general rule is that the agent tries to find biomass as close as possible to the heating plant in order to keep the transport distances and overall costs as low as possible. If two agents are longing for the same timber stand at a given raster cell, the agent visiting the cell earlier gets the available wood – the other one gets biomass that is left. Each agent searches until the yearly wood chip

demand for the associated heating plant has been collected. In addition, the transport distance - based on the generated transport distance raster - is monitored for each gathered m³ of timber and evaluated thereafter.

The evaluation of the transport distances as well as the pattern of collected timber over the given simulation period of several years shall give evidence for the spatio-temporal availability of biomass for energy purposes.

## 3. Agent-based Model for Collecting Wood for Heating Plants

The innovations of this model are not only the spatial resolution and the coupling of a GIS and an agent-based modelling environment using the Agent Analyst of ArcGIS (Johnston 2013) but also the subject matter which checks the hypothesis that the support of wood chips as fuel for heating plants is not a sustainable solution for renewable energy source projects in Central Europe. In this approach four specific procedures (indicated by the numbers 1 to 4 in Figure 1) were developed to model the universe of discourse accordingly.

The first procedure (# 1 in Figure 1) is a forest growth model based on empirical forest inventory data and a DEM, which represents the annual growth of the dominant two tree types (spruce and European beech) with respect to different altitude. This is done for all forest stands in Carinthia, which have an area of 580.000 ha of the complete 953.301 ha land area.

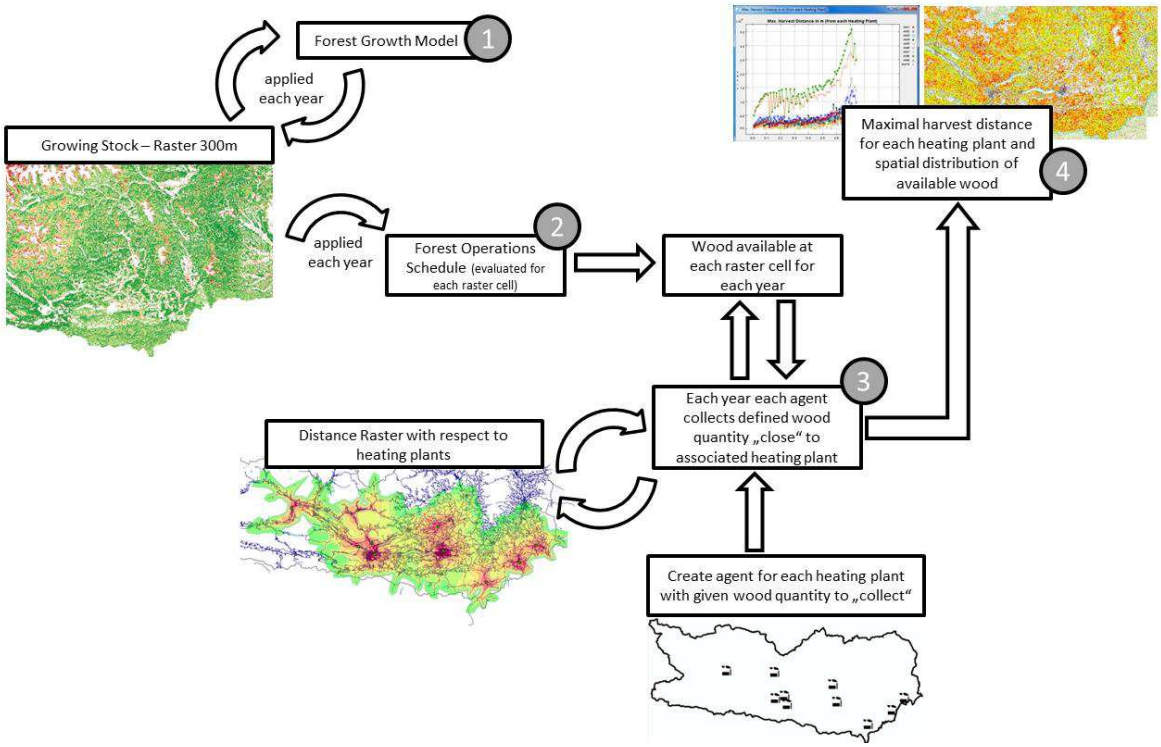



Figure 1. Approach to evaluate the wood chip availability with a given consumption of lumber by heating plants. The numbers in the circles define the sequence of operations carried out each year that is modelled.

The second procedure (# 2 in Figure 1) is an estimation model for the amount of available wood chips per raster cell (300m by 300m). Here the data of the forest inventory is essential to define a forest operations schedule, based on the standing timber and the "history" of the standing timber (i.e. prior forest operations, age of forest). In 2013 13% of the forest trees were under 20 years old, 60% between 20 and 100 years and 17% older than 100 years (10% were bushes etc.). The trees between 20 and 70 years of age were used for wood removal and that only every 10th year. So the forests in Carinthia are relatively old, which is similar all

over Central Europe. This fact and the decreasing maintenance of the forests are the main reasons for the decreasing supply of wood chips as fuel.

The third procedure (# 3 in Figure 1) is the simulation of the collection process for the prospector agents, which is based on a round trip concept to scan the landscape and collect wood chip material as needed by each heating plant. The prospectors collect wood in a systematic way as long as there is a demand of the supplied heating plant. When the collecting areas are overlapping, the prospector who is at a certain stand first, gets the wood. Further criteria for collecting the wood are a maximum of 20% removal of biomass per stand, 25% crop loss, removal only in stands having an age of 20 to 70 years and at an age of 100 years the stands are fully harvested and the amount of biomass is reset to zero.

The fourth and final procedure (#4 in Figure 1) is the calculation of different statistical parameters like the maximum harvest distance for each heating plant and of maps of the spatial distributions of available wood for each of the simulation years (Figure 2). Further ouput parameters of the simulation process are the amount of the total, the available and the not available wood for all of Carinthia (see Figure 3 Left) and the average age and the percentage of the harvestable forest stands (see Figure 3 Right). In these figures the proposition that the age of the stands is the main reason for a massive decrease in wood chips availability during the next decades is confirmed.

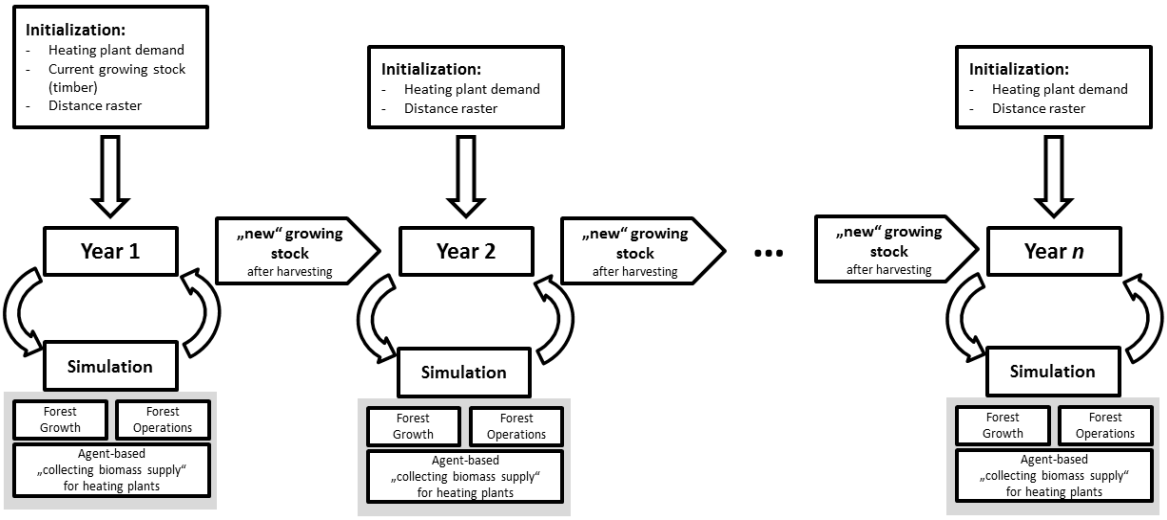



Figure 2. Agent-based simulation approach for each simulation year. This graphic depicts the temporal sequence of the simulation process and the necessary data for each yearly simulation.

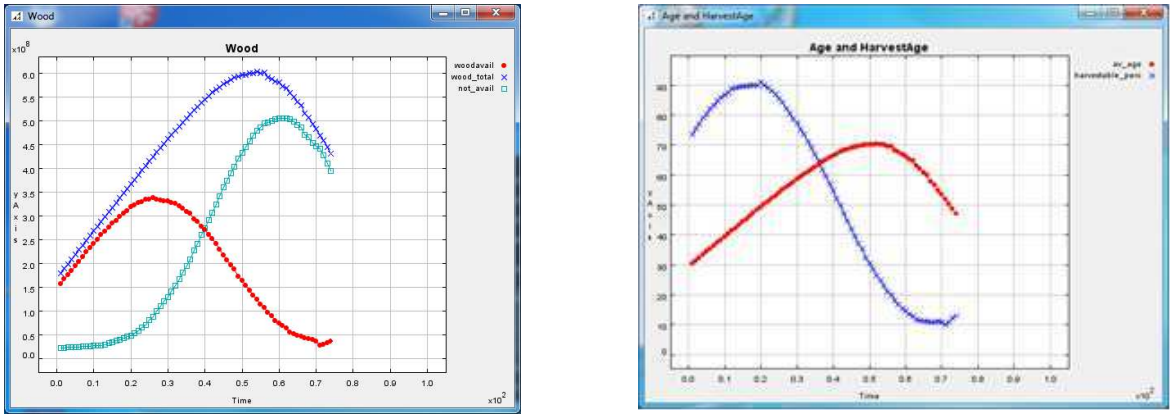



Figure 3. Results of the simulation process for Carinthia. Left. Amount of the total (blue), the available (red) and the not available wood (light blue), Right. Average age (red) and total percentage of the harvestable forest stands (blue).

## 4. First Results for Two Scenarios

The results of the spatial agent-based model are maps of the available amount of wood chips and statistics for the distance of timber haulage at each time step (from forest to heating plant). A prototypical implementation is applied to two scenarios to model the wood chip demand for heating plants in Carinthia. The first scenario models ten operational heating plants and the second one includes a fictional new heating plant for Klagenfurt.

The results of the agent-based spatio-temporal simulation processes for both scenarios are presented in Figure 4 and 5. Figure 4 shows the maximum transport distance from the forest to every heating plant of scenario 1 for each simulation year. The graph reveals that the maximum transport distance increases the longer the simulation runs. This is due to the fact that forests in the vicinity of a heating plant cannot supply the heating plants accordingly, and timber has to be transported over longer distances. This can also be justified by the increasing number of moves of each agent until the biomass supply for each heating plant is collected (see Figure 5). For scenario 2 the situation is similar to scenario 1 with an additional heating plant in Klagenfurt that increases the demand of fuel by a factor of 4.3.

The results of the scenarios indicate that the model is capable of modelling the timber usage of heating plants and the effects of heating plants on the "environment" – i.e. the forest, standing timber and transport distances – over a time frame of 50 years accordingly.

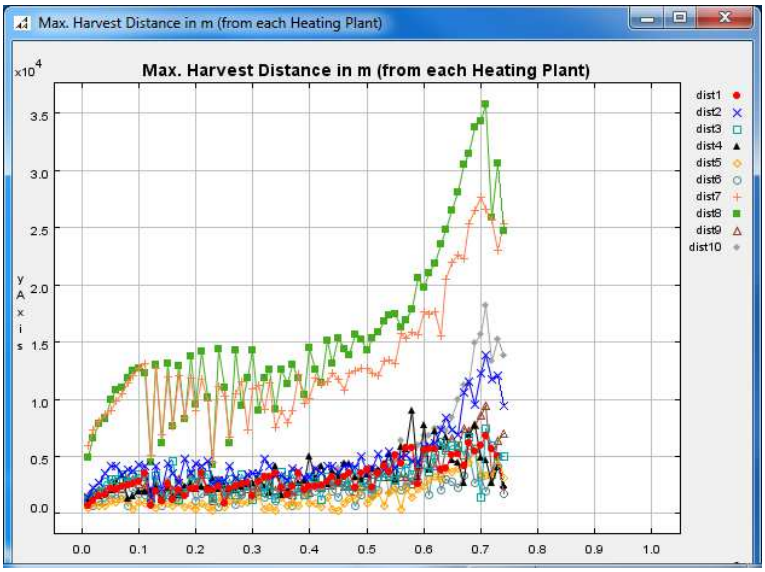



Figure 4. Results of the agent-based spatial simulation process for scenario 1. The results depict the transport distance to the last biomass collection point of each heating plant under review for each year of the simulation period (74 years for scenario 1).

## 5. Conclusion, Discussion and Future Work

The paper describes an approach to model the impact of wood chip heating plants on the availability of lumber for wood chip production. The model operates on a fine spatial and temporal granularity. In order to model the "consumption" of timber for heating purposes, an agent-based model coupled with a GIS is employed. The model is applied to two scenarios that include ten operational heating plants (scenario 1) and an additional high-demand heating plant located in Klagenfurt, Austria (scenario 2). The results reveal that the approach is capable of modelling the impacts of heating plants on forest and transport distances. Additionally, the results indicate, that, within the test area Carinthia one cannot fully rely on wood chips from local forests in order to fulfil the heating energy demand.

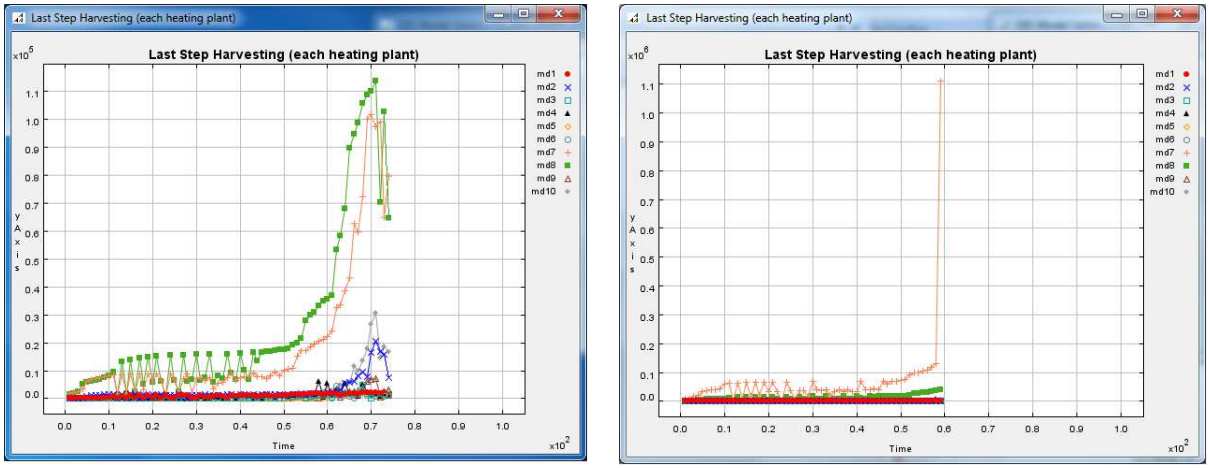



Figure 5. Results of the agent-based spatial simulation process for scenario 1 (left) and scenario 2 (right). The results depict the number of moves of each agent – i.e. prospector – until the timber demand for a heating plant for one year is collected.

In further projects the presented model can be used to simulate additional scenarios. The problems which will be worked on are the change of the types of forest operations and their timing, the influence of wood chip imports from other countries on the model results, the optimum of the number and the location of power stations which get along with the available amount of timber in Carinthia as well as the combination of the biomass power plants with other renewable energy sources. Furthermore the development of a simulation methodology supporting the Energy Master Plan for Carinthia (http://www.energie.ktn.gv.at/) is in progress.

## Acknowledgements

This study was supported by the Austrian Research Centre for Forests (BFW), which is greatly appreciated. In detail, the authors were given access to a general forest map, forest type and forest growth data for the province Carinthia. The authors would like to thank Mr. Bruno Regner of BFW for his efforts to support this research work.

## References

Arbeitsplattform Wald und Holz in Kärnten, 2007, Bilanz und Strategieplan über Aufkommen, Nutzen und Potentiale. Landwirtschaftskammer Kärnten, Klagenfurt.

Crooks A T and Heppenstall A J, 2011, Introduction to Agent-Based Modeling. In Heppenstall A, Crooks A T, See L M, Batty M (eds.), Agent Based Models of Geographical Systems, Dordrecht, Heidelberg, London, New York: Springer.

Johnston K M, 2013, Agent Analyst: Agent-Based Modeling in ArcGIS, Esri Press, Redlands. (available under http://resources.arcgis.com/en/help/agent-analyst/)

Mandl P, 2003, Multi-Agenten-Simulation und Raum – Spielwiese oder tragfähiger Modellierungsansatz in der Geographie? Klagenfurter Geographische Schriften, 23:5-34.

Möller B and Nielsen P S, 2007, Analysing transport costs of Danish forest wood chip resources by means of continuous cost surfaces. Biomass and Bioenergy, 31(5): 291-298.

Stampfer K and Kanzian C, 2006, Current state and development possibilities of wood chip supply chains in Austria. Croatian Journal of Forest Engineering, 27(2):135-145.

Van Berkel D B and Verburg P H, 2012, Combining exploratory scenarios and participatory backcasting: using an agent-based model in participatory policy design for a multi-functional landscape. Landscape ecology, 27(5):641-658.

# The Development of a Community and Platform in Support of Japanese OpenGeoData: A Case Study of the Urban Data Challenge of Tokyo 2013

T. Seto[1], Y. Sekimoto[2]

[1] Center for Spatial Information Science, the University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan, 153-8505
Email: tosseto@csis.u-tokyo.ac.jp

[2] Institute of Industrial Science, the University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan, 153-8505
Email: sekimoto@iis.u-tokyo.ac.jp

## 1. Introduction

Since the end of the last decade, the use of open data (secondary use and machine-readable formats) has emerged as a political and cultural movement for the realization of citizen participation. In particular, open government, citizen participation, transparency in government, and public and private cooperation were established as goals in the U.S. by the Obama administration in 2009 (Goldstein and Dyson, 2013; Morozov, 2013). In addition, in the "G8 Open Data Charter," which was declared at the G8 Lough Erne Summit in June 2013, geospatial information was specified as a kind of high-value dataset (Sui, 2014).

In Japan, the Fukui Prefecture's Sabae City began operations with Japanese open data at the end of 2010. After this date, the open data of government agencies developed rapidly, leading to the opening of Data.go.jp in December 2013; however, the data set of text and numbers of the White Paper are limited and do not contain completely open geospatial information. As of April 2014, the city published open data on the local governments of 38 cities (Figure1), but the information is not expressed in the house map; the location information of the infrastructure has also been published.

In contrast, we held a workshop (Urban Data Challenge Tokyo 2013: UDCT2013), the purpose of which was to promote participatory workshops with a data set that has been offered free of charge or with OpenGeoData from local governments. Along with the discussion of regional issues in local government, this data set can help to solve urban problems. In this paper, we report on the findings of UDCT2013 in examining the challenges and advantage of the OpenGeoData in Japan.

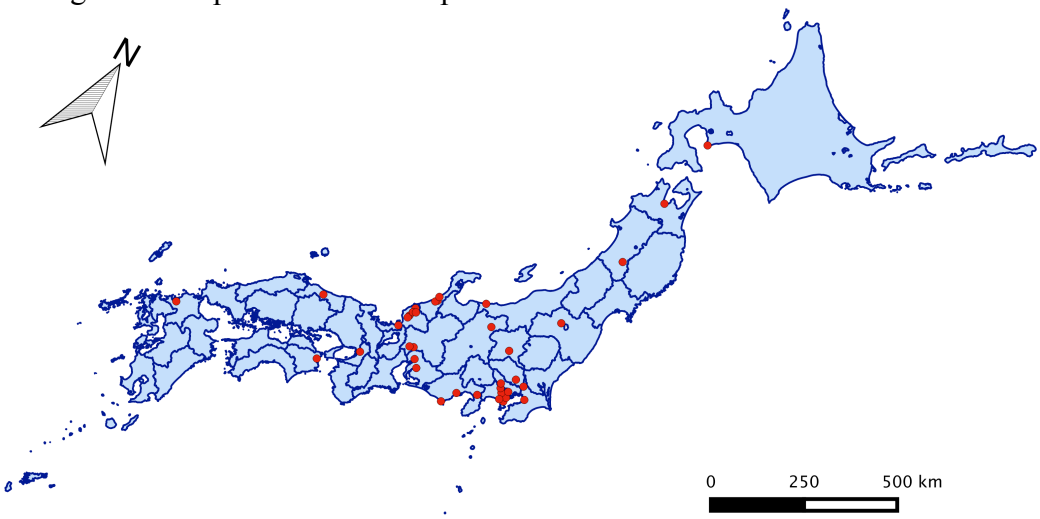



Figure 1: Japanese Open Data Cities (07/01/2014)

# 2. Practice and Community Development of the UDCT2013

## 2.1 Project Overview

In the UDCT2013, about 500 people participated in five events that were held throughout the year. The first event, which was divided into 10 teams, involved a free discussion of "what are the urban problems in our town." Many of the barriers to open data release were not clearly identified. After the discussion, a theme with 46 issues, including nine categories (population problems, town planning, infrastructure, town security, disaster prevention, transportation, agriculture, education, and tourism), was presented. The ideathon discussed the role that general citizens play in finding solutions. In addition to the results of the discussion, the disincentive to open public data became clear: "the type of governmental geographic data has been determined by guessing," "institutions are not equipped to handle open data," and the "accuracy of the original data is a concern."

Moreover, we collected data from the local government of the Tokyo metropolitan area; we obtained use permits from 21 cities (including the five-city data published by the CC-BY). In addition, we received assistance from the NTT Geospace Corporation and the Ministry of Land, Infrastructure, Transport and Tourism, which included geographic information on the infrastructure of the capital area as a whole. Thus, we have published nine types of 434 challenge datasets (Table 1.).

Table 1. UDCT2013 Challenge Datasets.

| Organizations /Categories | CC-BY Cities Datasets | UDCT2013 Joined Cities Datasets | Governmental/ Other Datasets | Total |
|---|---|---|---|---|
| | 5 | 16 | 3 | 21 |
| Population | 10 | 51 | 2 | 63 |
| Town Planning | 67 | 51 | 1 | 119 |
| Infrastructure | 17 | 22 | 6 | 45 |
| Security | 0 | 7 | 0 | 7 |
| Disaster | 27 | 58 | 2 | 87 |
| Transportation | 2 | 16 | 0 | 18 |
| Agriculture | 0 | 8 | 0 | 8 |
| Education | 9 | 37 | 0 | 46 |
| Tourism | 25 | 16 | 0 | 41 |
| Total | 157 | 266 | 11 | 434 |

## 2.2 Development of an OpenGeoData Portal with CKAN

We are building a portal site, which will allow workshop participants to use any OpenGeoData. Thus, with reference to GeoPlatform (Clark et al., 2013), we adopted the widely used CKAN. We opened the web site (UDCT-CKAN) in October 2013 as a portal on Geoserver, with overlapping geographic data (Figure 2). The datasets of UDCT2013 were collected from the target digital data and correspond to nine categories, mainly HTML (130 datasets), PDF (122 datasets), and image data, since information development on paper still flourishes in the local governments of Japan. Address information and facility statistical data are not tabular and are published only in PDF format. In addition, Google Maps API is used

in many cases; it does not deliver raw geodata (CSV, SHP, KML, and so on). This trend can be found in the open data of Japan, since CKAN does not fully support the display and exchange of the easy-use format of OpenGeoData. However, as open data is processed, they are translated into geodata, excluding exceptions, such as the Yokohama and Shizuoka prefectures that utilize GIS in some sections.

In the six months following the opening of the UDCT data portal site, about 18,000 users have accessed it and viewed 200,000 pages. About 6,000 users viewed only one page, the data lists, and the home page, but about 2,300 viewed more than ten pages, including specific datasets (Figure 3). In addition, the data related to the civil engineering infrastructure, such as railways and public facilities, have been viewed more than 100 times (datasets); the data on disaster management, such as tide levels and shelters, have also been consulted.

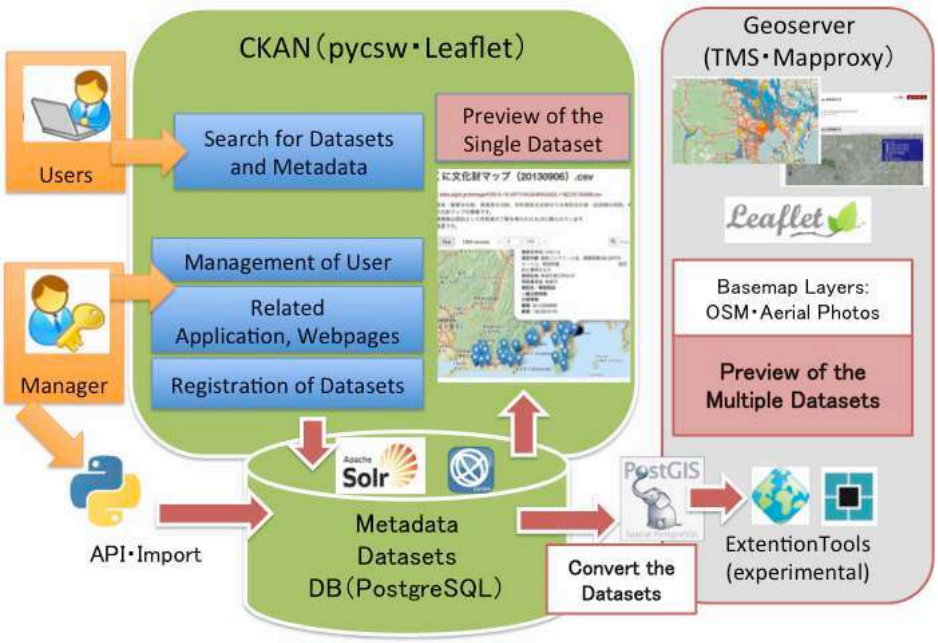



Figure 2: The System Configuration of UDCT2013 Data Portal Site.

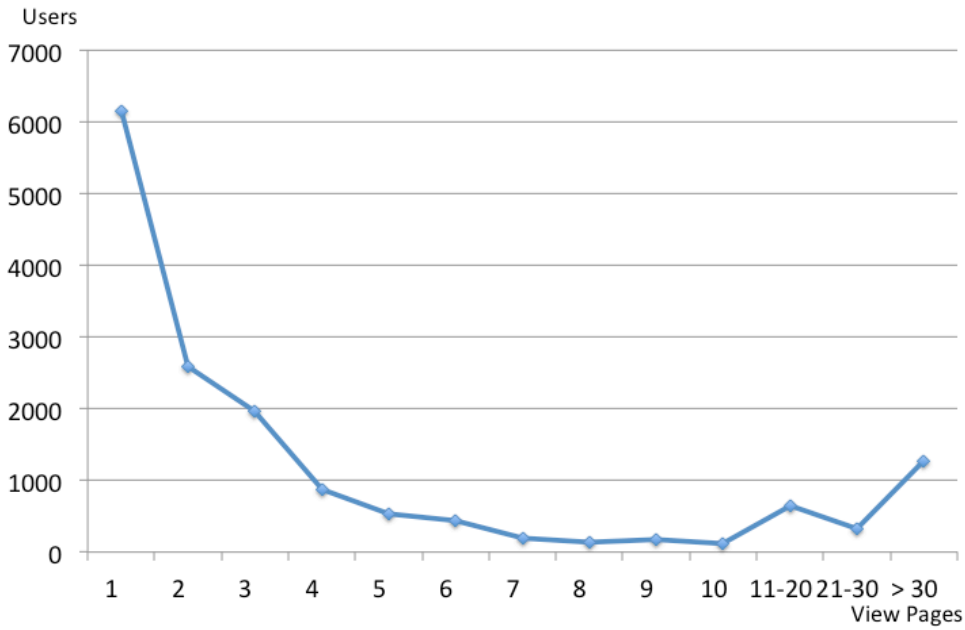



Figure 3: Page Views of the UDCT2013 Data Portal Site.

### 2.3 Application Development of Regional Issues with OpenGeoData

Seventy-five works were submitted through the UDCT2013 until the end of January 2014. The themes of the 75 work were, in order of importance, disaster prevention (17 works), community development (15 works), and tourism (10 works), but some combined several themes, such as transportation and population issues. However, the “application of the entry division” (31 works), “the data to complete Web services and software,” “the application and use of techniques for easy formatting,” “the data for the entry division” (16 works), and “ideas for the entry division” (28 works) were also presented (Table 2). In Japan, ideathons and hackathons for the promotion of data have increased rapidly in recent years, but the engineers and designers involved in the development using OpenGeoData are limited in number. However, application work has employed one type of dataset more than others.

A high level of awareness of disaster prevention and mitigation was triggered by the Great East Japan Earthquake. Further, the OpenGeoData of local governments, such those on shelters, fire hydrants, and hazard maps, have been published. It should be noted that much of the data that were submitted to UDCT2013in CKAN were converted into easy-to-use formats (KML or CSV), which promote the use of OpenGeoData.

Table 2. Applications Result of the UDCT2013.

| Category | Works | Average of using Datasets | Average of using Issues |
|---|---|---|---|
| Idea | 31 | 1.87 | 1.42 |
| Application | 28 | 3.68 | 1.71 |
| Data | 16 | 1.75 | 2.13 |
| | | | *p= 0.7768* |

## 3. Conclusions

As mentioned above, OpenGeoData from local government is becoming gradually available in Japan in recent years. In addition, OpenGeoData is beginning to be utilized in Japan to solve urban problems, such as disaster prevention and community development, through citizen participation. It is expected that local businesses, citizens, and engineers will use the application. We developed a platform that can be viewed on more than the web sites of local government when seeking solutions to regional problems and that takes advantage of geospatial information. It is easy to find and to download. In addition, real-time Geodata have become available in Japan, but the application of these data to regional problems is limited. Therefore, an increase in programs that combine enhancements and platforms that offer easy access to real-time GeoData is necessary.

## References

Clark R.J., Kuhmuench C., and Richard S.M, 2013, NGDS node deployment adoption of CKAN for domestic and international data deployment. *Proceedings, Thirty-eighth Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, California, 9 p.

Goldstein B. and Dyson L., 2013, *Beyond transparency: Open data and the future of civic innovation*. Code for America Press, San Francisco.

Morozov E., 2013, *To save everything, click here*. Allen Lane, London.

Sui D., 2014, Opportunities and impediments for Open GIS. *Transactions in GIS*, 18(1), 1-24.

# Pattern Analysis of Police Foot Patrol in Central London

Jianan Shen, Tao Cheng

Department of Civil, Environmental and Geomatic Engineering, University College London,
Gower Street, London WC1E 6 BT, UK

Email: {jianan.shen.13; tao.cheng}@ucl.ac.uk

## 1. Introduction

Understanding basic features of human movement is of great significance across many areas. The increasing availability of GPS-integrated equipments has enabled the researchers to look at the unprecedentedly abundant location history information. Previous researches have explored POI (Points of Interests) mining (e.g., Ashbrook, 2003; Cao, 2010), trajectory analysis (Giannotti, 2007) and mobility pattern generalization on normal pedestrians based on GPS records (Ye, 2009).

Despite its importance in policing activities, no research, to my certain knowledge, has been done for the analysis of police foot patrol footprints and their effects on crimes, partly due to the unavailability of such sensitive data. Police movements are unique in its purposive and dynamic nature. Unlike normal pedestrians who usually have only a few constant POIs (i.e., homes and working places), police officers often go on patrol with specific missions or shifting destinations influenced by emergency calls. Moreover, different officers may behave differently because of their different tasks and working habits. It is worthwhile to analysis the trajectories and mobility patterns of the police for the evaluation and improvement of their work and explore the nature of modern policing.

Here, we intend to present a framework of pattern analysis of police foot patrol, which is capable of 1) extracting police POIs from GPS log data; 2) summarising individual officers' traveling sequences and trajectories; 3) comparing similarity of trajectories and generalising mobility patterns of various groups of officers.

## 2. Data

### 2.1 Case study area

This study takes place in the Camden Borough (Figure 1), which lies to the north of central London. Five major police stations are located in this region, namely, West Hampstead, Hampstead, Kentish Town, Albany Street and Holborn. Research centred on the police pedestrian activities within Camden.

### 2.2 Data

The major data set captures officers' location stamps recorded by GPS-integrated radio sets portable on every officer in the field. The acquired data covering nearly 3 months (8 Dec 2011 - 29 Feb 2012) are exported from the Automatic Personnel Location Systems (APLS) of the Metropolitan Police. All of the 241525 records generated by 745 officers provide the information of call signs, device IDs, as well as the locations and times the record were made. Usually, the data is logged every 10 minutes, which is an acceptable temporal resolution for foot patrol. A record will also be inserted when an officer calls with his/her radio. However, when the radio is powered off or blocked for some reason, the logging will be stopped. One

call sign can only be used by one officer and will not be changed until s/he leaves his/her present unit. So we can consider that one call sign uniquely represents one police officer.

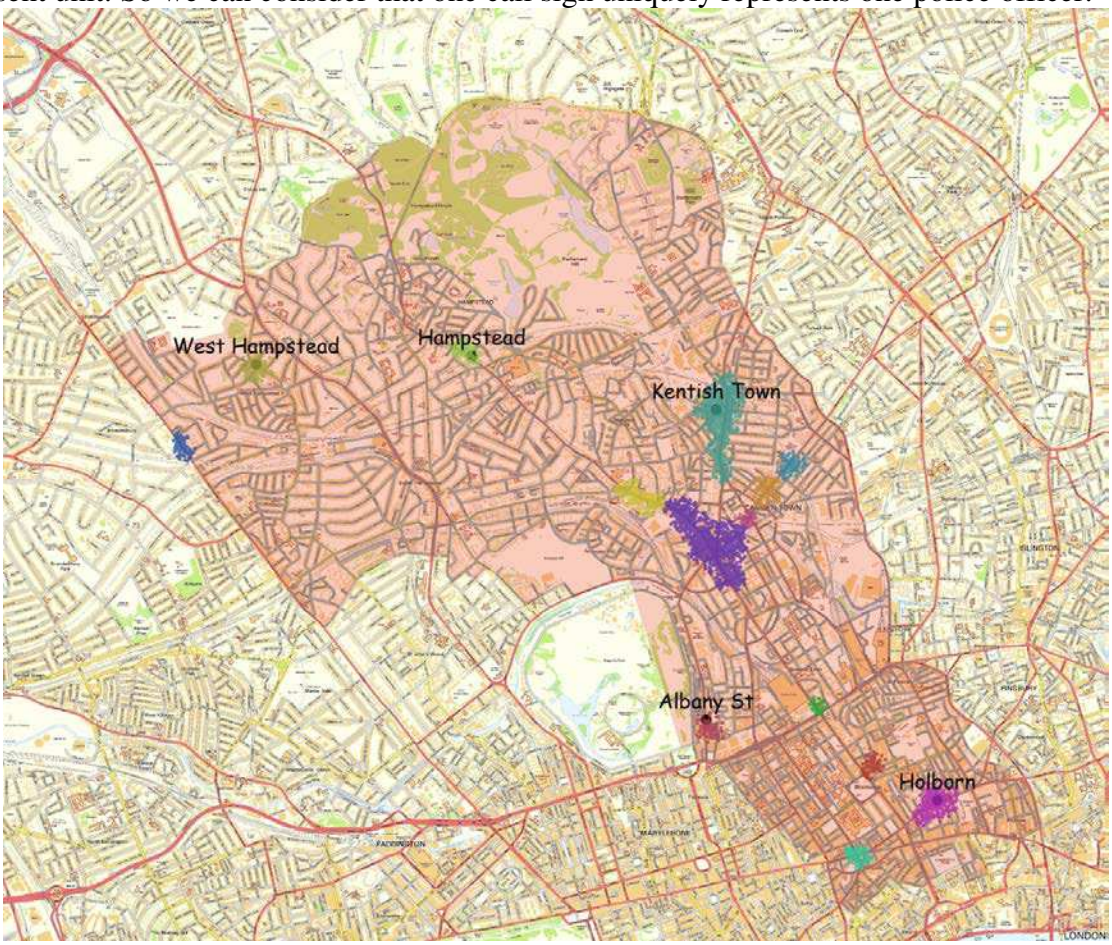



Figure 1: The borough of Camden and 15 POIs clustered in Feb, 2012

## 3. Methodology and results

### 3.1 Identifying POIs

Three major steps are executed for POI discovery, namely, events extraction and determination (Andrienko & Andrienko, 2011), density based aggregation (Zhou, 2004) and validation. For moving objects like patrolling officers, the locations where they stop are much more meaningful than where they simply pass by (Palma, 2008). Therefore, stay points (Figure 2) are identified as places where the offices stay still or wander through at a prominently low speed.

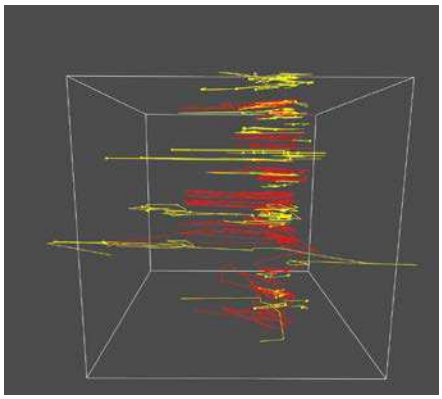


Figure 2: 3D visualization of real trajectories (polylines) and stay points (dots) of two officers (Officer 1 in Red and Officer 2 in Yellow)

The massive activities observed nearby police stations are common reflection of police daily routine and bear little information. To discover more semantically significant POIs, outliners and records nearby the police stations are removed before density based clustering. DBSCAN (Ester, 1996) is used for the clustering of stay points. Several schools, tube stations and major streets are discovered to be points of interests for the police. The clustering outcomes should be associated with land use and crime data for validation and further analysis. Here we use them for the trajectory analysis.

## 3.2 Trajectory analysis

Following the concept of stay points, an officer's trajectory can be considered as a set of stays and moves, where the clustered POIs are the most important parts of the trajectories. The spatial temporal trajectory is abstracted as the sequence in which the officer visits each POI and the time s/he arrives at and leaves the POIs (Li, 2008). This method can simplify the spatio-temporal analysis as shown in Figure 3 and lay down foundation for trajectory similarity analysis. In this way, the detailed walking routes are neglected, given the insignificance and poor sampling rate of GPS log. In similarity analysis, walking time intervals between POIs, staying times, quantity of common POIs and travelling sequences are compared and weighted to evaluate the synthetic similarity degree of trajectories.

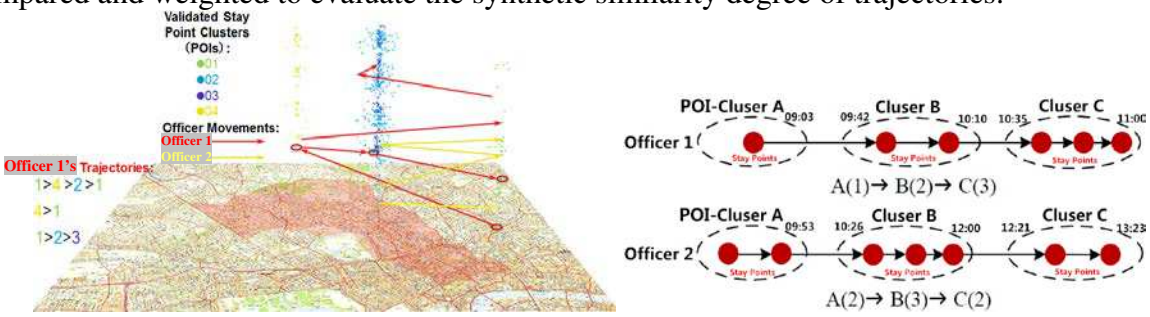



Figure 3: The extraction of trajectories, turning trajectories into series of POI visiting sequence

## 3.3 Mobility Pattern Analysis

Figure 4 demonstrates a temporal comparison of activities by showing both daily and weekly periodic patterns of different officers in a month. The identity information of officers are removed for confidential reasons. The colour blocks represent the number of records the active officers generated and uploaded with APLS in one hour. The temporal discrepancies clearly show that activeness of some officers are more intensive and constant than others. The temporal pattern of officer 4 shows that s/he is only on task in certain days and active for a continuous period of time. Officer 1 is on task in most days but his/her working time is unevenly distributed while the active time of officer 3 concentrates on afternoons every week.

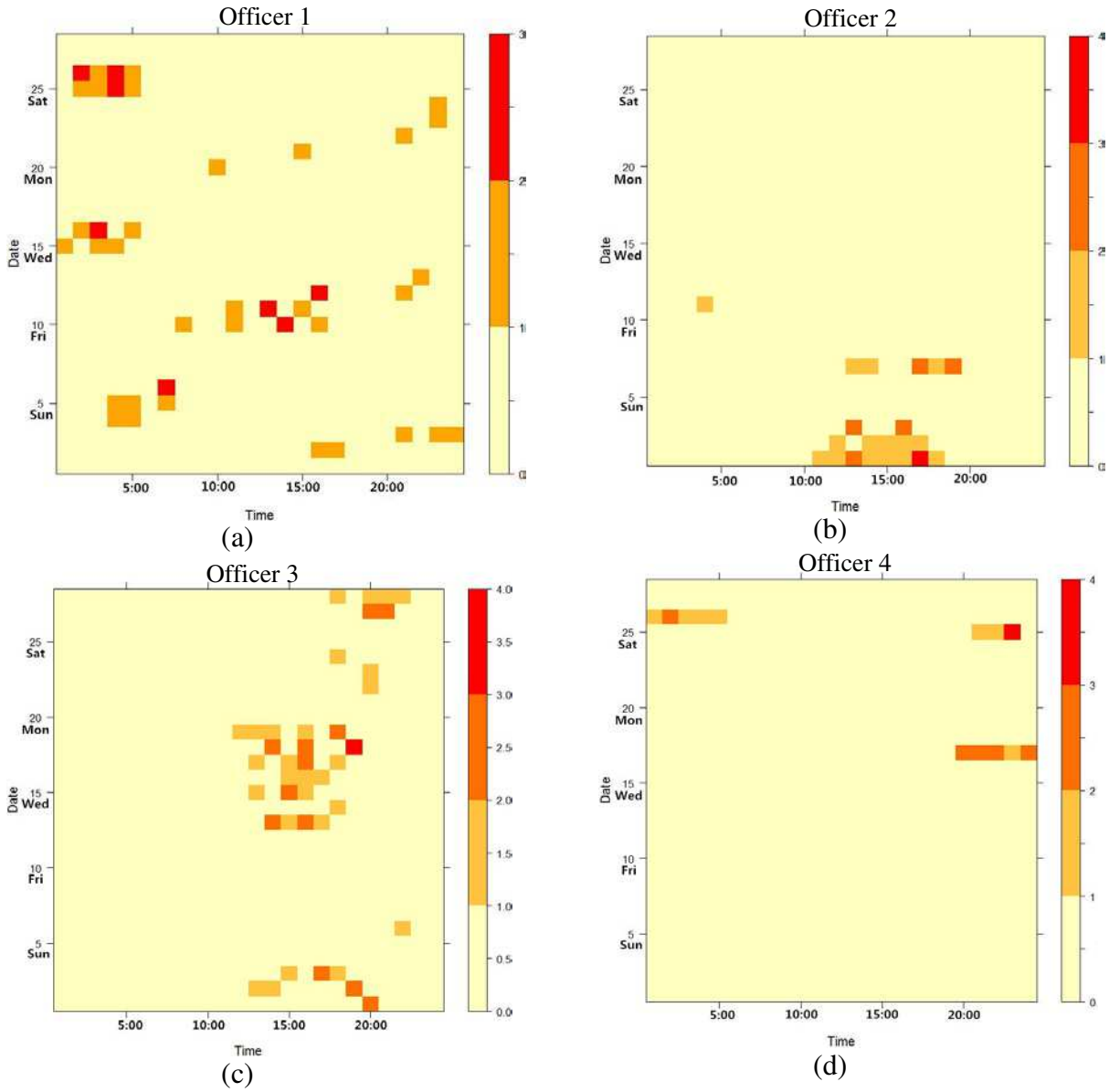



Figure 4: Heat map of patrol activeness in one sampled month
a) Officer 1; b) Officer 2;
c) Officer 3; d) Officer 4

Based upon the sequences of visiting POIs (see Figure 3), we are able to see the pattern discrepancies of different officers as showing in Figure 5. Some officers always stay at the same place in working time (such as officer 3) while other officers keep walking through various POIs (such as officer 1) because of their task differences. Here, officer 1 also share similar aggregation in cluster 1 and 3 with officer 1 but with lower intensity in other subordinately significant POIs. These may reveal potential interdependent relationship between the officers, especially in those areas with intensive works. The spatio-temporal combined analysis of patrol patterns is to be explored in future works.

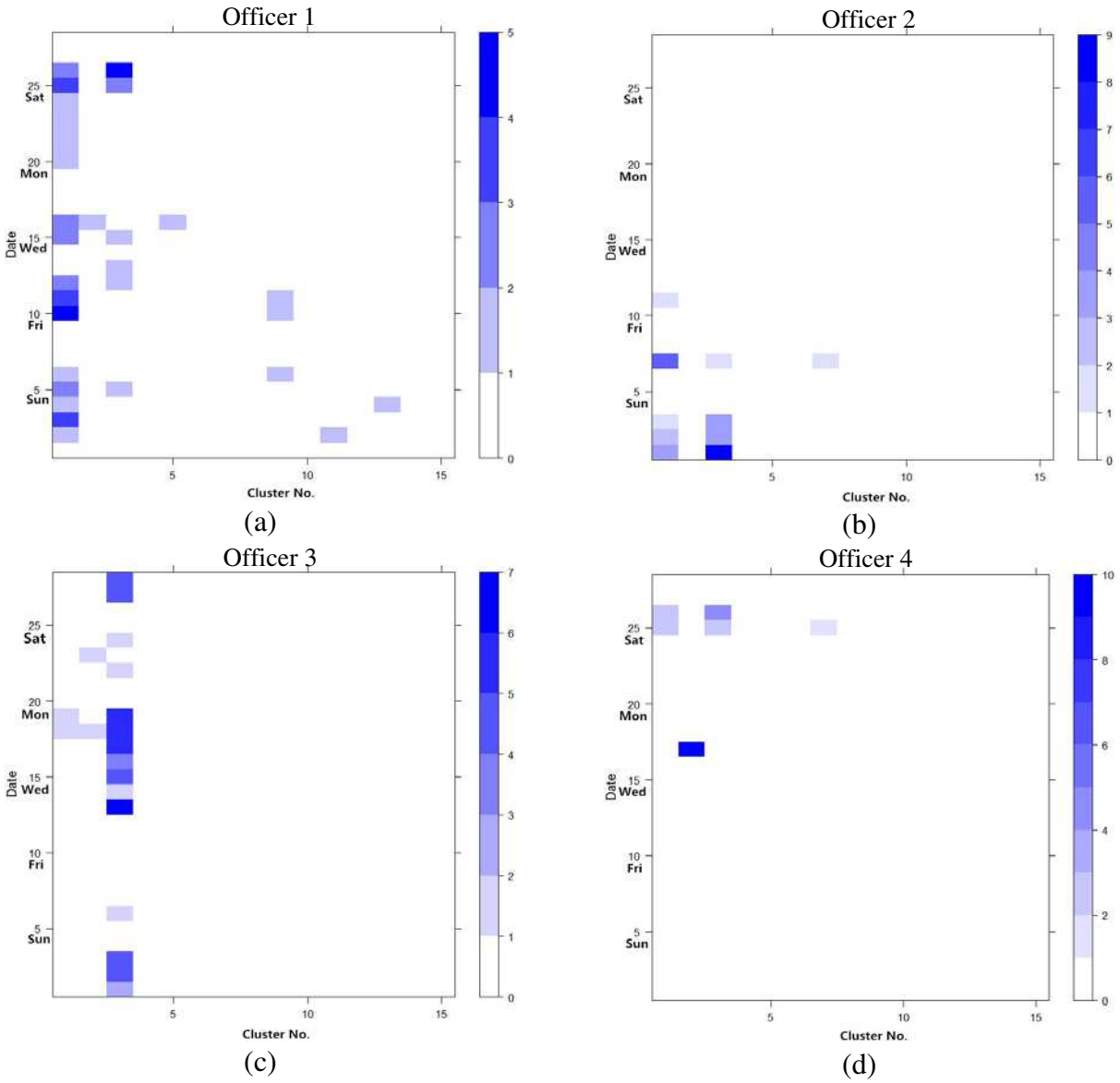



Figure 5: Heat map of visit frequency to different clusters (POIs) in one sampled month
a) Officer 1; b) Officer 2;
c) Officer 3; d) Officer 4

## 4. Summary and future work

There are papers that have looked at general police presence as a whole and tested the influences in an overall statistical scale (Ratcliffe & Taniguchi, 2011). However, it is a critical omission that the detailed process of foot patrol has never been empirically studied. The Camden APLS data enabled the study that others cannot proceed with due to the lack of modern GPS-enabled police and authorization issues. In this research, DBSCAN algorithm is used to search for POIs; Sequence analysis took the advantage of the temporal features of patrols; Similarity analysis of trajectories revealed general movements share by group of officers and laid down the foundation for mobility pattern generalization.

Further works may include improvement of clustering and validation, trajectory similarity analysis and spatio-temporal patrol patterns generalisation based on hierarchical clustering. After the similarity evaluation algorithm is established, similarity thresholds and trajectory models will be set based on Damerau-Levenshtein distance and longest common subsequence

to cluster different types of footprints so that universal patrol pattern can be generalised. K-means clustering and KNN classification will be used and compared for behaviour labelling and evaluation. Besides, the data will further be associated with other dataset such as crimes and emergency calls. The calculation process will be parallelized to optimize the time consumption.

## Acknowledgements

This work is part of the project - Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks (www.ucl.ac.uk/cpc), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1). The data provided by Metropolitan Police Service (London) is highly appreciated. Jianan Shen would like to acknowledge the China Scholarship Council for his PhD research.

## References

Andrienko, G. and Andrienko, N., 2011, From Movement Tracks through Events to Places: Extracting and Characterizing Significant Places from Mobility Data. Symposium on Visual Analytics Science and Technology, October 23 - 28, Providence, USA. 161-170.

Ashbrook, D. and Starner, T., 2003, Using GPS to Learn Significant Locations and Predict Movement across Multiple Users. Personal and Ubiquitous Computing, Atlanta, USA. 275-286.

Cao, X., 2010, Mining Significant Semantic Locations From GPS Data. VLDB Endowment, Singapore. 1009-1020.

Ester, M., 1996, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD, Portland, USA. 226-231.

Giannotti, F., 2007, Trajectory Pattern Mining. KDD, San Jose, USA. 330-339.

Li, Q., 2008, Mining User Similarity Based on Location History. Geographic Information Science, November 5-7, Irvine, USA.

Palma, A., 2008, A Clustering based Approach for Discovering Interesting Places in Trajectories. Symposium on Applied Computing, March 16-20, Fortaleza, Brazil. 863-868.

Ratcliffe, H., Taniguchi, T., 2011, The Philadelphia Foot Patrol Experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots. Criminology, Philadelphia, USA. 795-831.

Ye, Y., 2009, Mining Individual Life Pattern Based on Location History. Mobile Data Management, Taipei, Taiwan

Zhou, C., 2004, Discovering Personal Gazetteers: An Interactive Clustering Approach. Geographic Information Science, November 12–13, Washington DC, USA. 266-273.

# The Use of VGI for Spatial Interaction Modelling

K. Sila-Nowicka[1], T. Oshan[1], J.Vandrol[1], A.S. Fotheringham[2]

[1]Centre for GeoInformatics
Department of Geography & Sustainable Development
University of St Andrews
Irvine Building, North Street, St Andrews, Fife, UK
KY16 9AL
Email: ksn2@st-andrews.ac.uk

[2]School of Geographical Sciences and Urban Planning
Arizona State University
Coor Hall, 5th floor
975 S. Myrtle Ave.
Tempe, AZ 85287

## 1. Introduction

Spatial Interaction is a term for any type of movement over space that results from a decision-making process. A crucial idea of spatial interaction is that there is a transfer of a physical entity, such as a good or a person, between an origin and a destination. More specifically, examples of spatial interaction include, commuting journeys, shopping trips, recreational visits, travelling to school, telephone calls, and various forms of social networking. Consequently, spatial interaction is arguably one of the most significant topics within human geography because it incorporates so many important types of human behavior (Fotheringham & O'Kelly, 1989). Through understanding spatial interaction we can ask questions such as "where is the ideal location for a new supermarket?" or "how will the location of a new store affect demand at existing stores". Explosive increases in the amount of available geo-referenced data, which can be easier and cheaper to collect than conventional surveys, is making it possible to consider new ways of deriving observed interaction flows. These can then be used in order to calibrate models. One particular example is Volunteered Geographic Information (VGI) (Goodchild, 2007), in the form of trajectories collected by GPS loggers, which is high in temporal resolution and collected at the individual user level. Before flows can be extracted, the data must be filtered, processed and broken into segments which can serve as coherent inputs into models. This research aims to develop a framework for processing GPS trajectories into flows between pre-defined origins and destinations and then to subsequently test how well the derived flows can complement traditional survey data in the calibration of spatial interaction models.

## 2. Methods and applications

### 2.1 Data and case study

The process of collecting and understanding human mobility patterns is becoming increasingly important for research, policy makers and the private sector. As we continue to generate massive amounts of spatially referenced data, an explicit focus on GPS traces, personal paths, human mobility-based behaviour and spatial interaction is essential. To verify if VGI, in the form of GPS trajectories, may be used for spatial interaction modelling, GPS trajectories from volunteers were collected.

The selected study region (Figure 1) covered the three largest towns in the Kingdom of Fife in Great Britain (Dunfermline, Kirkcaldy, Glenrothes), each of which has about 40 thousand inhabitants. Medium size towns were selected in order to provide a new source of information about mobility patterns in an environment other than a metropolis, such as Beijing (Zheng, et al., 2010) or London. Advertisements for the GPS survey were sent out and the selected volunteers were asked to carry a passive GPS device for a week. Overall, 206 volunteers participated in the survey which resulted in more than 4 million points, each with a longitude, latitude and temporal coordinate ( $p_i(\varphi, \lambda, t)$ ) at about 5 second intervals, providing an average of 26 hours of data per participant.

The collected raw GPS trajectories were first filtered to facilitate further analysis. Filtering directly removes some elements from a GPS dataset. Different filters may be applied for different purposes, such as deleting points associated with a weak satellite signal (Hightower, et al., 2005), identifying non-movement within a trajectory (Subramanya, et al., 2006) (Ashbrook & Starner, 2002) (Hightower, et al., 2005) or by using timestamps to remove trips which were too short (Froehlich & Krumm., 2008). We applied all of these filtering techniques to maintain the highest quality of GPS trajectories seeking to minimize outlier points which could affect subsequent analysis.

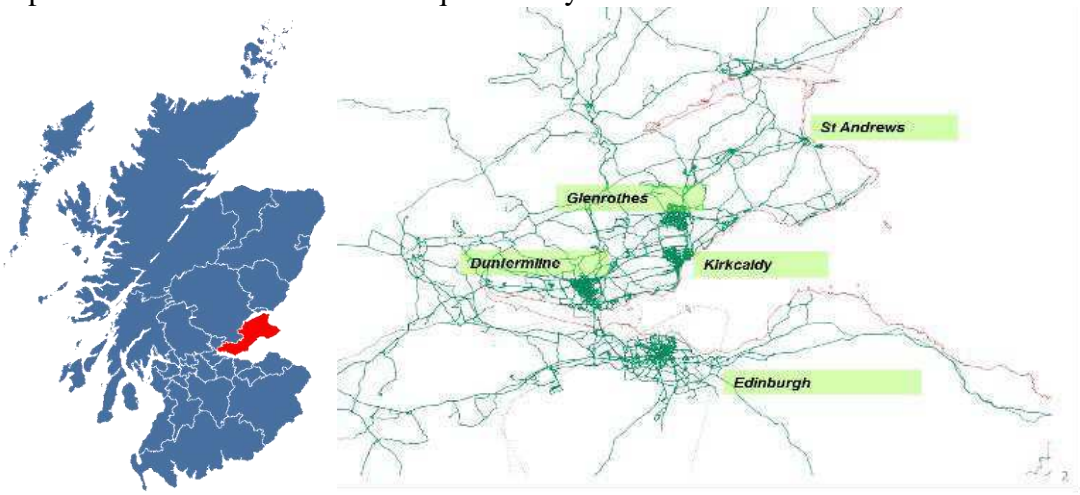



Figure 1: Region of interest presenting 3 largest towns in Kingdom of Fife, Great Britain with the base map in form of collected trajectories.

## 2.2 Origins and destinations extraction

Individual trips were then extracted from each participant's overall daily movements. This is carried out by first defining stops within a trajectory which delineate the beginning and end of separate journeys. Trajectory segmentation has been approached by utilizing spatiotemporal characteristics of collected points (Buchin M, et al., 2011), hidden Markov models (Bui H, et al., 2001) or Bayesian networks and is often compared against contextual information from sources such as Open Street Map to assess the accuracy of their predictions (Liao, et al., 2007). In this research we leverage the spatio-temporal distribution of points associated with trajectories to develop a segmentation algorithm. To create an origin-destination matrix for spatial interaction modelling, we needed to first cut the user-collected trajectories into smaller segments which represent changes in transportation mode. The starts and ends of these smaller segments are treated as origins and destinations and then classified as Points of Interest. Points of Interest (POIs) were both inferred from trajectories and compiled from existing sources such as the Ordnance Survey database. Primary POI's used

for this study include individual places of residence, which were verified by geocoding addresses supplied by participants, places of work, and major shopping or leisure centres. Places of residence or work they were aggregated to administrative units to prevent privacy issues. By combining trips to common origins and destinations the data can be encoded in an origin-destination matrix and used as input to calibrate a spatial interaction model.

### 2.3 Spatial Interaction modelling

Data derived from the volunteered GPS trajectories were tested in three different modeling contexts. First, all trips defined as commutes between residences and places of work were used to calibrate a doubly-constrained spatial interaction model utilizing Scottish data-zone municipal boundaries to represent origins and destinations. Next, a production-constrained model was employed in order to analyze the attractiveness of retail shops which allowed us to comment on the spatial decisions associated with shopping behavior when travelling from different data-zones to key shopping outlets and calibrate models to recreate the reality (Figure 2).

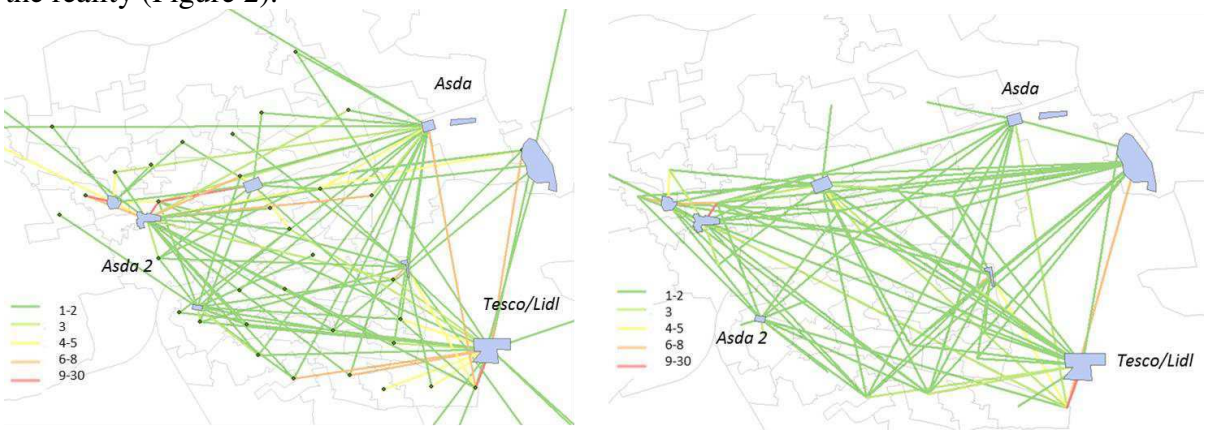



Figure 2: Real (left) and estimated (right) shopping trips from centroids of data zones in Dunfermline

Finally, trips associated with leisure activities such as jogging, playing golf or exercising at gyms, were used to calibrate a production-constrained model and a destination-constrained model which seek to understand the attractiveness of leisure centers and the characteristics of those engaging in leisure activity, respectively. For all of the models, the appropriate parameter estimates can be reported, along with their associated standard errors, to assess the quality of the estimates as well as the standardized root mean square error statistic to test uncertainty of the model.

## 3. Discussion

Disadvantages of using GPS trajectories as a source of spatial interaction include GPS inaccuracy and the time-consuming process of collecting, cleaning and processing data. Nevertheless, the enormously amounts of GPS data which can be collected about daily movements provides a unique alternative to existing National Travel Survey data, which are updated every 10 years. It is therefore worth investigating new possible data sources and how they can be manipulated to understand the processes by which people make spatial decisions.

Our focus is mainly on data processing and POI extraction, and how the uncertainty associated with this new source of data affects its reliability when used within spatial analysis. A final outcome will be to determine what additional information should be included in the process of data collection to improve the reliability and usefulness of raw

GPS trajectories. Volunteered GPS trajectories are a promising source of information for uncovering patterns in human behaviour and spatial reasoning and this study can be seen as stepping-stone to extend spatial interaction modelling into the wider realm of volunteered geographic information and dis-aggregated movement data.

## Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme - Marie Curie Actions, Initial Training Network GEOCROWD under grant agreement n° FP7-PEOPLE-2010-ITN-.

## References

.

Ashbrook, D. & Starner, T., 2002. Learning significant locations and predicting user movement with GPS. s.l., Wearable Computers, 2002

Buchin M, Driemel A, van Kreveld M, Sacristán V ,2011, Segmenting trajectories: a framework and algorithms using spatiotemporal criteria. J Spatial Inf Sci, 3

Bui H, Venkatesh G, West G, 2001, Tracking and surveillance in wide-area spatial environments using the abstract hidden markov model. International Journal of Pattern Recognition and Artificial Intelligence, 15

Fotheringham, A.S. & O'Kelly, M.E., 1989, Spatial Interaction Model: Formulations and Applications. Kluwer Academic Publishers, London

Froehlich, J. & Krumm., J., 2008. Route prediction from trip observations.. s.l., SAE.

Goodchild, M. F., 2007. Citizens as sensors: The world of volunteered geography. GeoJournal 69 (4), p. 211–221.

Hightower, J. et al., 2005. Learning and recognizing places we go. s.l., UbiCOMP 2005.

Liao L, Fox D, Kautz H (2007) Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. International Journal of Robotics Research, (26)

Subramanya, A., Raj, A., Bilmes, J. & Fox, D., 2006. Recognizing activities and spatial context using wearable sensors. Citeseer, Proc. of the Conference on Uncertainty in Artificial Intelligence.

Yang Yuea, Lia, Y.-g. & Yehd, A. G., 2012. Exploratory calibration of a spatial interaction model using taxi GPS trajectories. Computers, Environment and Urban Systems, 36(2), pp. 140-153

.

# A Basic Study on the Region Partitioning based on Semi-Supervised Classification

A. Takizawa[1]

[1]Osaka City University, Sugimotocho 3-3-138, Sumiyoshi-ku, Osaka, Japan
Email: takizawa@arch.eng.osaka-cu.ac.jp

## 1. Introduction

When small-scale space such as a building or open space is deigned, designers seem to design them expecting that some events or activities of users might occur there. Some events are intended by the designer, and some ones are not intended. Predicting the occurrence of such events is important. Applying a classification model to handle statistically these problems is one method for the problem. If we apply a classification model to such problems, we have to distinguish the region in which events occurred or not, and label them differently like Positive or Negative, respectively. Since the area of the region where events occur is generally smaller than that of region without any event, the classification accuracy will be low if only the event occurrence point is labeled with Positive. Since spatial attributes tend to have spatial autocorrelation nature, there might be some rationality to regard the neighbor of an actual event point as potentially event occurrence region. However, it is difficult to determine such region preliminarily because of the ambiguity of space. Moreover, in recent years, spatial designs that the relationship between space and function is ambiguous are increasing and this makes the analysis more difficult.

In this study, we propose a novel spatial analysis method for relatively small space. This method employs the semi-supervised learning (Zhu and Goldberg 2009) concept and aims to classify and explain the difference between regions on event occurrence in high precision. This method can also automatically obtain the region in which events tend to occur. Then we test this method with some artificial data.

## 2. Framework of the proposed method

### 2.1 Problem setting

In this study, we perform the analysis on the two dimensional plane. The plane is discretized into cells whose length of a side is $l$ in order to simplify handling the shape of the region described in the next section. Let $c \in C$ denote a cell and its set, respectively. The problem is to classify each cell into one of two classes on the event occurrence. The class labels are P that denotes event occurrence and N that denotes event non-occurrence. Let $o_c \in \{\mathrm{P}, \mathrm{N}\}$ denote the objective variable of cell $c$ and $\boldsymbol{v}_c = (v_c^1, \ldots, v_c^N)$ denote the vector of explanatory variables of $c$ where $N$ is the number of the explanatory variables. Then, the data of cell $c$ is represented as a set of tuple $(\boldsymbol{v}_c, o_c)$. In the original data set, P is labeled only to the cell in which events occurred, and other cells are not labeled yet. Let $e \in E \subset C$ denote the event cell in which an event occurred and its set and $R_e \in \Re \subset C$ denote a neighbor cell within the radius of $r$ cells from $e$ and its set (neighbor region). We allow labeling P to a cell of $R_e$. Other cells $c' \in C \setminus \Re$ are labeled with N (see Figure 1). Let $C_\mathrm{P}$ and $C_\mathrm{N}$ denote sets of cells labeled with P and N, respectively. The whole area is divided into two areas that do not overlap each other, that is $C = C_P \cup C_N, C_P \cap C_N = \emptyset$. Let $dv(C, \Re)$ denote the function that

partitions $C$ into two areas with $\mathfrak{R}$, $f(\cdot)$ denote a classifier and $err(\cdot)$ denote the function that returns a classification error.

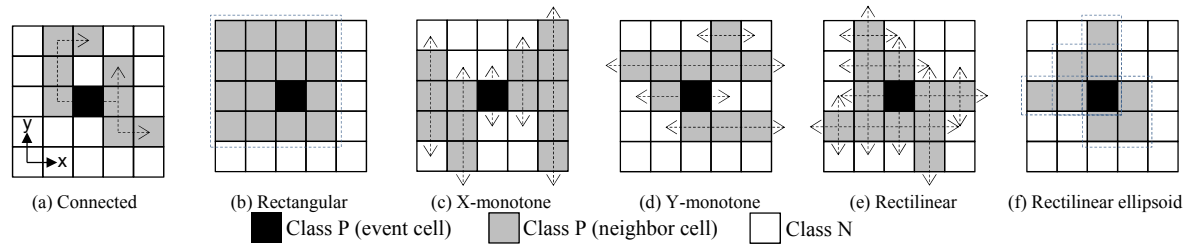



Figure 1: Distinction of cells on event occurrence and examples of the region family. Dashed lines represent the characteristics of each region.

Finally, we formulate the problem that finds the optimal $\mathfrak{R}$ for minimizing the classification error as follows

$$\min_{\mathfrak{R} \subset C} \ err\left(f\left(dv(C, \mathfrak{R})\right)\right), \quad \text{s.t. A region constraint for } R_e \in \mathfrak{R}. \tag{1}$$

The region constraint is described in the next section.

## 2.2 Region family on a discrete plane

We regard that two cells are connected and located in the same region if they share a same edge. Figure 1 illustrates some examples of the region family on a discrete plane. (a) is a simple connected region that satisfies the definition but this region lacks in unity. Moreover, the combination of those possible patterns is $\mathbf{O}(2^r)$ and this becomes too large as the increase of the radius. (b) is a rectangular region but its freedom of the shape seems to be too little for our analysis. (c) and (d) are called a x(y)-monotone region that continuously intersects a straight line parallel to the x(y)-axis, respectively. (e) is called a rectilinear region that is x-monotone and y-monotone region, has a unity and some degree of complexity on the shape. (f) is called rectilinear ellipsoid region (Chen et al. 2004) that is composed of rectangles one of whose vertices is the event cell. The inclusion relation of the above regions is (b) $\subset$ (f) $\subset$ (e) $\subset$ (c), (d) $\subset$ (a).Among them, rectilinear ellipsoid is useful for our analysis since the visibility in the region from the event cell is guaranteed.

## 2.3 Evaluation of the classification accuracy

There are many kinds of the definition of classification errors. Table 1 is the confusion matrix that is typically used for calculating such errors. From this table we can define $Accuracy = (TP + TN)/(TP + FP + FN + TN)$ , $TPrate$ (Sensitivity) $= TP/(TP + FN)$ and $TNrate$ (Specificity) $= TN/(TN + FP)$. If the distribution of a particular class in a dataset is biased, we should use *Sensitivity* instead of *Accuracy*.

Table 1. Confusion matrix.

| | | Predicted Positive | Negative |
|---|---|---|---|
| Actual | Positive | *TP* (Num. of corresponding data) | *FN* |
| | Negative | *FP* | *TN* |

### 2.4 Selection of a classifier and solver

Our method can use any classifiers and the optimization problem (1) is approximately solved with a meta-heuristics. It might be possible to solve the problem (1) exactly by mixed integer problem if we use a linear discriminant function that is mathematically simple. Though, if we use it, the good classification result cannot be expected since the performance of the classifier itself is not high.

## 3. Implementation

We employ rectilinear ellipsoid as a region and balanced error rate (BER) as an error function. BER considers evenly both accuracies of positive and negative classes, that is $BER = 1 - (TPrate + TNrate)/2$. We also employ classification by associate emerging patterns (CAEP, Dong et al. 1999) as a classifier. In the previous study (Takizawa 2013), we used CAEP for classifying the place of street crimes and got good result. Population-Based Incremental Learning (PBIL, Baluja 1994) is used for optimizing the shape of the region. In order to apply PBIL we define the probability vector and code the parameter of the shape of rectilinear ellipsoid regions. Regions are coded for every event cell independently even if their regions are overlapped each other. Figure 2 illustrates the correspondence of the code and shape where an event cell is located at the center in the case of radius $r = 3$. The numbers around the cells represent the index of the probability vector and gene vector of PBIL. The values of the probability vector represent the probability that the value of the corresponding gene vector becomes 1. The shape of a region is parameterized as the height of the upper and lower cells for each x-coordinate. Four heights are possible for the cell indexed with 1 as illustrated in Figure 3. Rectangles whose opposite corners are the cell with the encoded height and the event cell respectively are drawn and the shape of the region is determined.

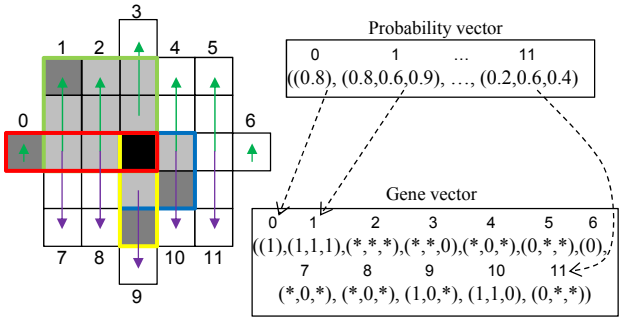



Figure 2: Gene coding for a region.

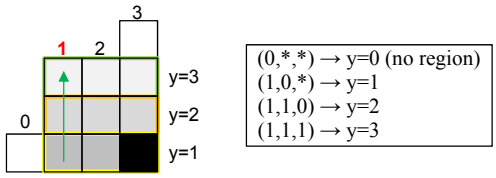


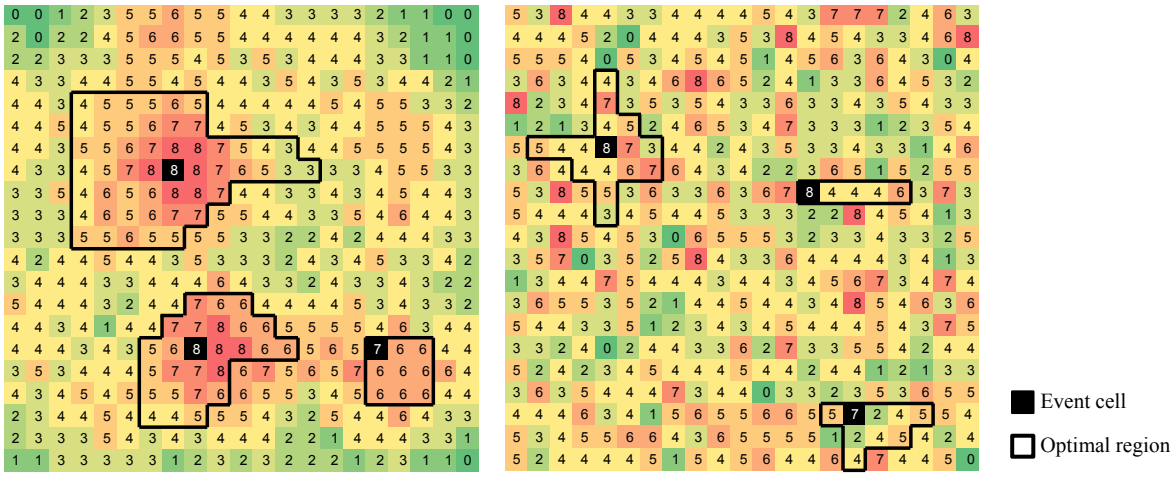

Figure 3: Decoding the gene information for the cell indexed with 1.

## 4. Case study

We perform numerical experiments by using artificial data. The target area is square whose length of one side is 21 cells. Four explanatory variables, that are independent each other and take a value of 1, 2, or 3, are considered. The value of each variable is spatially distributed like two-dimensional Gaussian distribution. Then the values of four variables are summed up and three cells whose total is more than seven are selected as event cells. Let this target area be Case 1 (see Figure 4(a)). Moreover, cells of this area are randomly rearranged and let this area be Case 2 (see Figure 4(b)) in order to evaluate classification performance depend on the presence or absence of spatial auto correlation. The maximal radius $r = 6$.

Table 2 lists the classification accuracy. The accuracy of Case 1 is better than that of Case 2. TPrate is higher than TNrate in both cases. The optimal region of Case 1 tends to be wider

than that of Case 2. The shapes of regions are not so simple and it implies that the high degree of freedom of rectilinear eclipse is appropriate for modelling such region.




(a) Case 1 (b) Case 2

Figure 4: Target area and optimized result. Colour gradation and numbers represent the sum of the value of explanatory variables.

Table 2. Classification accuracies.

| | *BER* | *Accuracy* | *TPrate* | *TNrate* |
|---|---|---|---|---|
| Case 1 | 0.088 | 0.866 | 0.988 | 0.837 |
| Case 2 | 0.264 | 0.680 | 0.800 | 0.672 |

## 4. Conclusion

In this study, we proposed a novel spatial analysis method for relatively small space such as an architecture and street level. Then we tested this method with some artificial data. We got the conclusion that this method can demonstrate relatively good classification performance for the spatial data that has some spatial autocorrelation.

## Acknowledgements

This study was supported by a Grant-in-Aid for Scientific Research (C) (No.25420633).

## References

Baluja S. 1994, Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning. *Technical Report, Pittsburgh, PA: Carnegie Mellon University*, CMU-CS-94-163.

Chen D., Chun J., Katoh N. and Tokuyama T., 2004, Efficient Algorithms for Approximating a Multi-dimensional Voxel Terrain by a Unimodal Terrain, *Proceedings of 10th Computing and Combinatorics of Conference (COCOON2004)*, LNCS 3106, 238-248.

Dong G., Zhang X., Wong L. and Li J., 1999, CAEP: Classification by aggregating emerging patterns, *Proceedings of the 2nd International Conference on Discovery Science*, 30-42.

Takizawa, A., 2013, Emerging Pattern Based Street Crime Analysis - Street Level Spatial Analysis of Crime Location Associated with Built Environment in Fushimi Ward, Kyoto City -, *Journal of Architecture and Planning (Transactions of AIJ)*, 78(686), 957-967.

Zhu X. and Goldberg A., 2009, *Introduction to Semi-supervised Learning*, Morgan & Claypool Publishers.

# GPU Accelerated Chart Visualization In GIS Using Point Splatting

Matthias Thöny[1], Renato Pajarola[1]

[1]Visualization and Multimedia Lab, University of Zurich
Email: mthoeny@ifi.uzh.ch, pajarola@ifi.uzh.ch

## 1. Introduction

Interactive visualizations are a key aspect for modern geographic information systems. Domain users as well as all other groups of users have the need to interactively explore and analyse large datasets. An important aspect of geographic visualizations is to display statistical representations in a geographically referenced context such as election data or aspects of population geography. The basic tools to visualize statistical information are chart diagrams such as pie or bar charts. Recent hardware developments make it possible to use GPU based techniques to speed up visualization programs. GIS programs already profit in many different areas from GPU acceleration. Examples are complex mathematical calculations, such as coordinate transformations or image processing techniques as well as more possibilities for interactive tasks such as editing geographic information. In this extended abstract we present a showcase for integrating different types of charts within a perspective 3D environment in real-time by using modern GPU acceleration techniques.

## 2. Visualization of charts using point splatting

Point splatting is a well-known technique to visualize point based data sets. Point splatting can be used very efficient for point information as shown in Gross et al.(2007), Kobbelt et al. (2004) or Sainz et al.(2004). The key idea is to display points by deforming each point shape in a way that no holes are visible in the resulting image. The technique indicates, that there is no complex mesh structure generated and therefore the main advantage of point splatting is the reduced geometric complexity. This visualization technique fits perfectly the needs for large-scale point dataset, e.g. models from laser scanning.

The original point splatting technique uses a textured rectangle for every visible point containing the pixel information of the point splat. The textured rectangle can be stretched or otherwise adjusted based on the camera's viewing angle to cover bigger or smaller regions with one single splat image. Modern graphics hardware makes use of this concept when drawing points and provides hardware acceleration within the rendering pipeline. Figure 1 shows the semantic structure of a colored point within the point-rendering pipeline. A point is defined as location in a local 3D space of the application. When accessing the rendering pipeline the points are mapped to a so-called camera space. In this camera space, the points are rendered as squared areas, according to the given point size. This point size specifies the point shape extent in pixel. A fragment shader program is performed for every fragment of this area. Within this fragment shader program it is possible to access the fragment coordinate relative to the screen and also the point coordinate relative to the point shape. This point coordinate is aligned between 0.0 and 1.0 within the point splat. The circle shape in normal point rendering appears, because the fragment program discards all fragments with a certain distance to the center point. These fragment programs are hardware parallelized, so that the rendering can take advantage of this multithreaded technique.

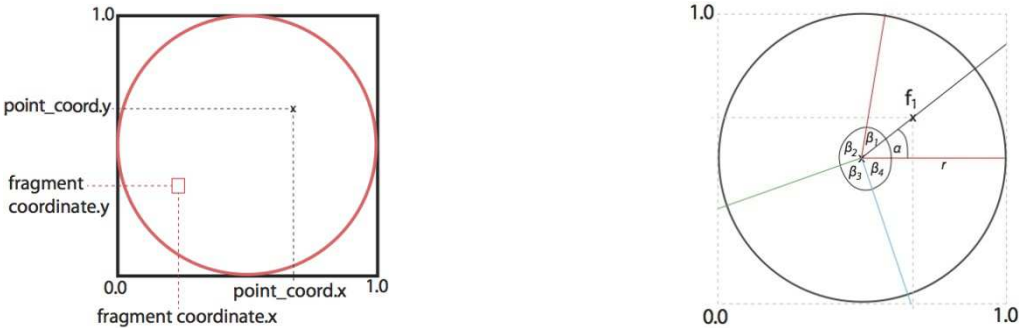



Figure 1: a) Structure of a point splat area. b) Sketch of a pie chart with four sections.

It is possible to use the hardware accelerated point splatting method for the purpose of information visualization by extend the fragment shader programs. Figure 2 shows example results from our application using point splatting for charts. We implemented three types of charts: bar charts, pie charts, and rose diagrams.

A division of the x-axis according to the number of incoming data table columns computes the bar charts. The fragment shader is executed for every fragment and decides about the correct color and group for every fragment $f_i$. Depending on the amount of columns of the incoming data table, the bar sections can be identified by a modulo-n-check of each x-coordinate of every fragment $f_i$. The fragments color can be chosen based on a predefined color lookup table. Fragments exceeding a bar section are discarded to make fragments not belonging to bars opaque.

Pie charts can be computed using a division of angle relations. Figure 1b) shows a sketch of the concept. The resulting sections correspond to the data table columns. For every fragment $f_i$ within a point splat, the corresponding angle $\alpha_i$ is computed. This angle $\alpha_i$ is compared against every angle $\beta_j$, which contains one certain data column extent. $\beta_j$ are computed from the input data values and represent relative percentages to $\sum_{j=0}^{n} \beta_j$ for all $j$ between 0 and the number of table columns. According to bar charts, the color value is chosen by the index $j$ from a color lookup table. In addition, a fragment is discarded if the distance to the center is bigger than the splat radius $r$ to produce a round shape.

Rose diagrams are a combination of pie and bar charts. In case of a rose diagrams, pie sections have the same angle size, but each slice differs in extent. This means, that the modulo-n-check is done based on the angle, so that the section or group can be identified for every fragment $f_i$. In addition the radius of each section is adjusted to relative values. The example in Figure 2 shows a red line within the rose diagram showing the 50% border, which means for our example that the absolute majority is reached.

## 3. Result and conclusion

Our prototype geographic visualisation software is able to show several thousands of charts as well as scene objects, such as terrain data within a interactive 3D environment. The system allows interactive changing of diagram shapes as well as selection of individual highlights within the charts. The approach of using point splats has two major advantages. First, rendering is slightly faster, because the geometric objects are simple point locations containing attribute data and not rectangles containing textures. Second, the memory consumption on the graphic card is reduced. The charts act like normal point objects regarding their geometric behavior. This means, that overlaps can be handled and spatial algorithms such as level-of-detail approaches could be applied.

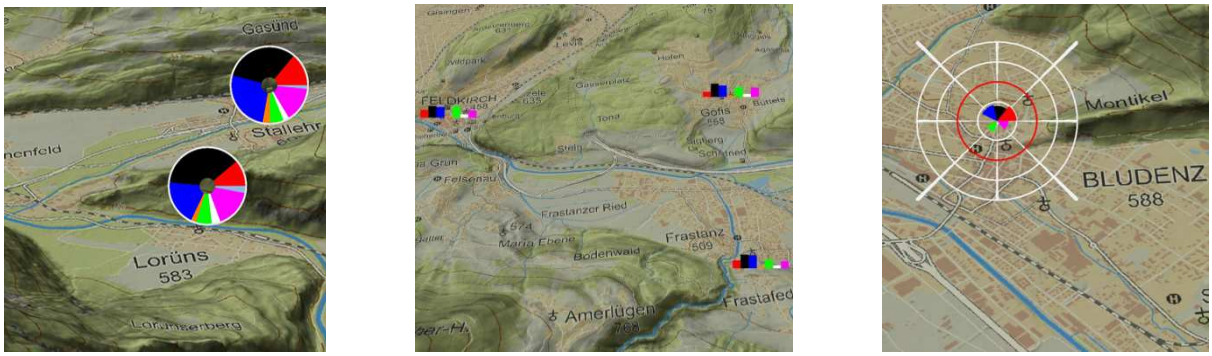



Figure 2: Austrian National Election data in 2013 displayed with 3 different types of charts: a) pie charts or donut charts. b) Bar charts. c) Rose diagrams.

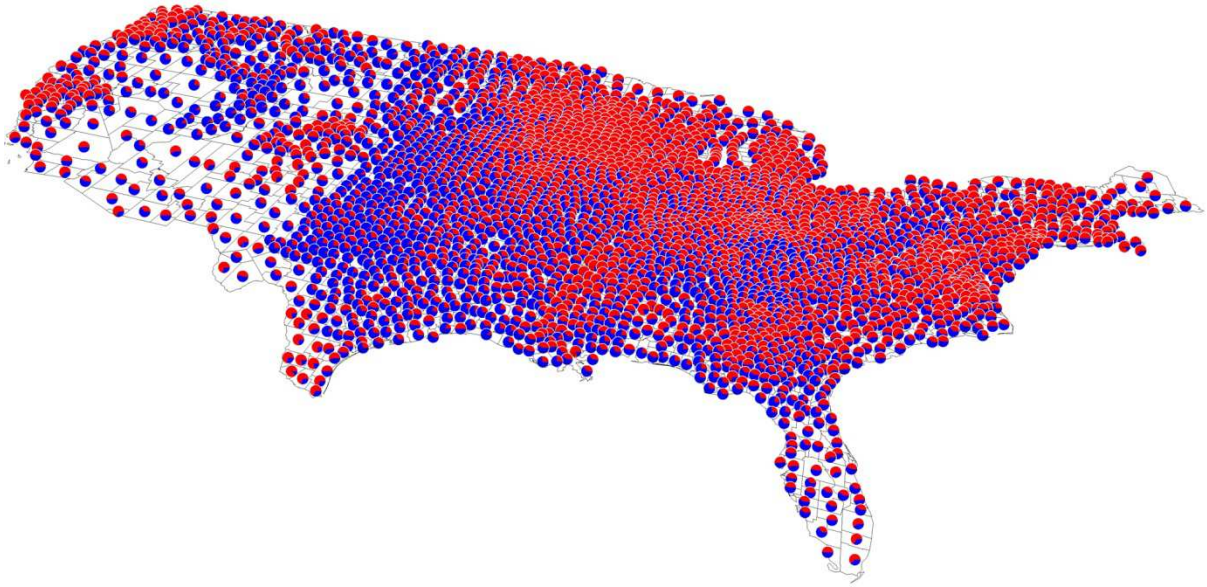


Figure 3: US Presidential election results of 2012 based on per-county information.

Figure 3 shows a dataset containing 4588 charts. This data set can be explored interactively in 3D with several hundred frames per second on a desktop computer. Mobile devices or other portable graphic devices can be targeted by this speed up. Most of the time, these devices are limited in graphics memory but with increasing parallelism of the graphics chips, easy parallelisable methods like point splatting will have increased importance. In the future we plan to further explore algorithms for automatically rearranging and adjustment of data elements based on occlusion information as well as further integration of level of detail algorithms for large scale geographic visualizations.

## Acknowledgements

We would like to thank the Landesvermessungsamt Feldkirch for providing the Vorarlberg data sets as well as the Austrian National Election data of 2013 for the region Vorarlberg.

## References

Gross M., Pfister H. Editors, 2007, *Point-Based Graphics*, Series in Computer Graphics, Morgan Kaufmann Publishers.

Kobbelt L., Botsch M., 2004, A survey of point-based techniques in computer graphics. *Computers & Graphics* 28(6), 801-814.

Sainz M., Pajarola R., Lario R, 2004, Points reloaded: Point-based rendering revisited. *In Proceedings Eurographics/IEEE VGTC Symposium*, 121-128.

# Harmonizing Level of Detail in OpenStreetMap Based Maps

G. Touya [1], M. Baley[1]

[1]COGIT – IGN France, 73 avenue de Paris 94165 Saint-Mandé France
Email: guillaume.touya{at}ign.fr

## 1. Introduction

As OpenStreetMap (OSM) is growing larger every day, practical applications based on OSM data are flourishing, but the initial goal of the project was to produce open topographical maps. A quick look at the default map output provided by OSM shows that it is difficult to create good legible maps out of the huge amount of data in OSM. One of the mains obstacles to the creation of good legible maps from OSM data is the heterogeneity of its level of detail (LoD). This heterogeneity is troublesome for large scale maps, the focus of this paper, as undetailed objects are often inconsistent with the detailed features of the map (Touya 2012). The understanding of a map is highly dependent on the way the reader grasps spatial relation between map objects. As a consequence, LoD inconsistencies are mainly damaging when occurring between spatially related objects (Figure 1).

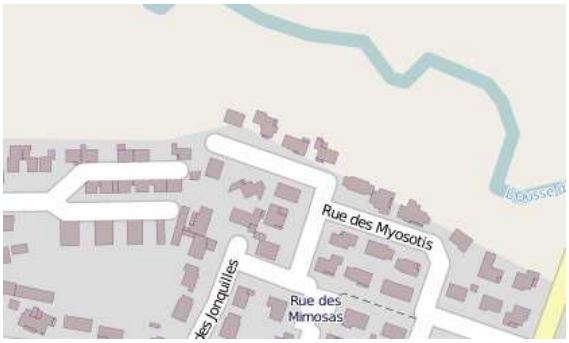



Figure 1: Buildings should be inside the built-up area to make it a readable spatial relation.

In order to remove the LoD inconsistencies, the preliminary step is to infer the level of detail of each OSM object. This is not the focus of the paper, so we consider that this LoD has already been inferred using the method from (Touya & Brando 2013).

Dealing with LoD inconsistencies in large scale maps requires the transformation of undetailed objects to make them consistent with detailed objects when objects share a spatial relation that helps understanding the map. We call such a process harmonizing LoD, implying that the harmonization increases LoD or at least preserves it, while map generalization (Sester et al. 2014) decreases LoD by simplifying the detailed objects. The automation of harmonization raises two questions. Is it possible to automatically harmonize OSM maps? Is it meaningful to transform data without any information on ground truth to make it more detailed? The ongoing work presented in this paper seeks to explore both questions by experimenting first attempts of automatic harmonization on OSM data.

In the second section presents several automatic harmonization methods for cases identified in the OSM dataset. The third part shows some experimental results and the fifth part draws conclusions and explores further research.

# 2. Harmonization Operations for OpenStreetMap Data

In this section, three cases of LoD inconsistencies are highlighted and methods are proposed to achieve harmonization in such cases.

## 2.1 Complete Aggregations

Geographical datasets often comprise high level objects that are aggregates of lower level objects of the dataset: a city is an aggregate of buildings, roads and parks for instance. In OpenStreetMap, such aggregate objects are very common and are generally less detailed than their components. This generates the most frequent LoD inconsistencies with components that lie just outside the aggregate (Figure 1). Harmonizing such inconsistencies consist in extending the aggregate to include the objects that are obvious components of the aggregate. Figure 2 shows the three steps of the proposed harmonization algorithm: (i) buffers are computed around the components to include, (ii) buffers are merged to the aggregate, and (iii) the outline is simplified to get a consistent resolution all along, as far as possible.

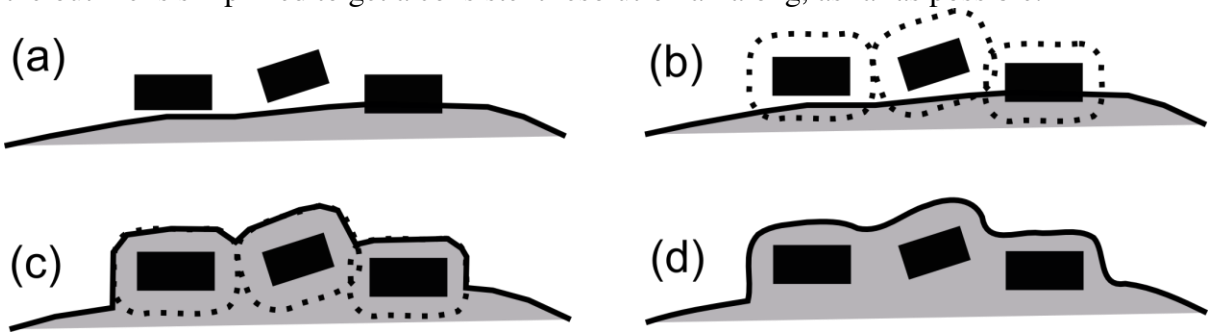



Figure 2: Steps to extend built-up areas. (a) Initial state (b) compute buffers of buildings (c) merge the built-up area with the buffers (d) simplify to preserve resolution.

## 2.2 Relocate

Relocation deals with inconsistencies where the positional accuracy is the most significant aspect of the lack of detail, i.e. some undetailed objects are clearly misplaced regarding some detailed objects. Misplaced objects should be displaced in the map to improve its readability. For instance, trees are often poorly detailed because hard to capture precisely and tree alignments often overlap road symbols. The algorithm proposed to relocate the trees removes the overlap and forces the alignment (Figure 3). Tree alignments on the right and on the left of a road are first identified. Then, right and left offset lines are computed, on which trees are projected to be aligned and off the road symbol. When a tree is at a crossroad, the projected position is the intersection of both offset lines.


Figure 3: Inaccurate trees overlap a road – harmonization aligns them on a road offset.

### 2.3 Re-Delimit

Re-delimitation is the modification of an undetailed object, with a more detailed outline, using the detailed objects in relation to figure out the more detailed outline. For instance, a detailed cycle way cannot intersect the outline of an undetailed lake captured on satellite images. The cycle way has a certain width, so the harmonized lake outline cannot be adjacent to the cycle way, a gap between the cycle way and the lake must be added.

Figure 5 shows that, sometimes, paths crossing a lake are just bridges. To avoid bad harmonization in such cases, the re-delimitation algorithm is improved with a pre-step that automatically identifies parts of a path that belong to a bridge. Two characteristics of bridge sections are used for the identification:

– The middle of a bridge is more “inside” the lake than its extremities (Figure 4a and b),
– The angle between the bridge and the nearest lake shore is close to 90° (Figure 4c and d).

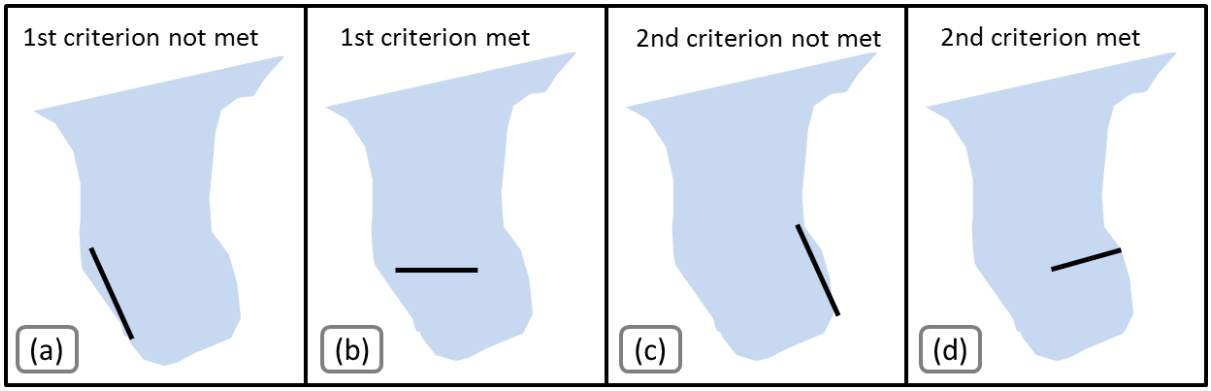



Figure 4: Criteria to characterize bridge segments.

## 3. Experiments

All presented algorithms have been implemented on the open source software GeOxygene (Bucher et al. 2012). Several datasets were extracted all over the world, but mostly in France. French datasets are useful as they can easily be compared to the authoritative maps we have access to, produced by IGN, the French mapping agency. Figure 5 shows some results of the lake re-delimitation algorithm.


Figure 5: (a) a cycle way intersecting a lake (b) harmonized lake outline (c) case with bridges: bad harmonization (d) harmonization with automatic detection of bridges.

Experiments on large datasets confirmed our assumption on the difficulty to find the best parameter values for harmonization algorithms. For instance, there is no obvious value for defining how far a building can be to be considered as “just outside” a built-up area. Harmonization algorithms parameters are context-dependent: parameters values are adapted to some situations and other situations require different parameter values. For instance, Figure 6a clearing uses the standard set of parameters empirically defined, but does not look like the real clearing drawn in the IGN map (Figure 6c). A specific set of parameters, which fails for most other cases, gives, here, much better results. Consequently, a knowledge base will be required to achieve the automation of harmonization in a complete map.

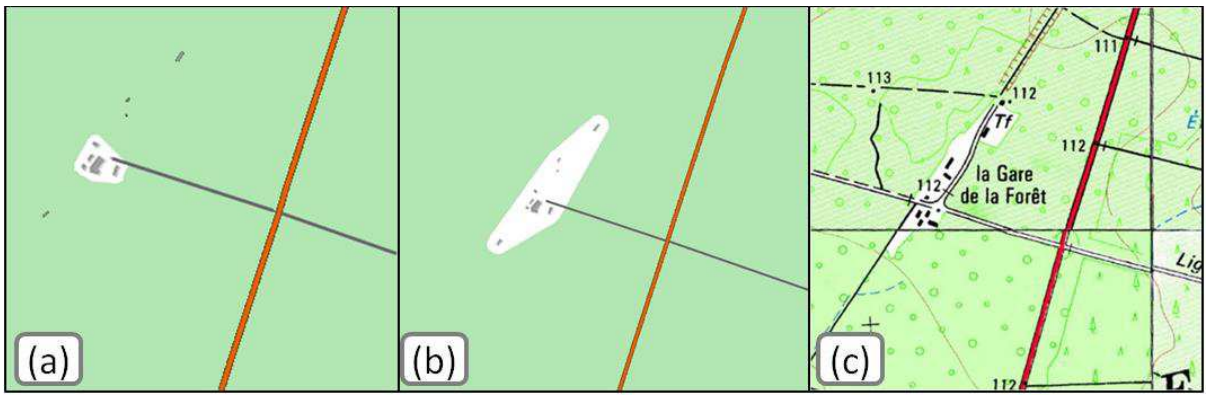



Figure 6: The clearing created with the standard parameters (a) does not look like the clearing in the IGN map (c); different parameter values give closer results (b).

A first basic evaluation was carried out by comparison with detailed authoritative datasets. A manual matching helped comparing the harmonized features with their authoritative counterpart. Shape and positional similarity measures show that the harmonized features are

closer to “reality” than the initial OSM features. Further quantitative evaluation is planned to confirm the first assumptions made on these experiments.

## 4. Conclusion and Further Work

To conclude, this paper introduces a new cartography problem raised by OpenStreetMap data, namely LoD harmonization, which improves the LoD of some undetailed objects in large scale maps to remove inconsistencies. Several types of harmonization operations are proposed and experimented on OSM datasets.

Further research should clearly focus on harmonization processes, to be able to automatically chain harmonization operations using a knowledge base, and solve complex problems that involve many objects like generalization or conflation processes (Harrie and Sarjakoski 2002, Touya et al. 2013). Furthermore, as harmonization operations transform data into something realistic but false, evaluation methods should be improved and user tests should be investigated, to know if map user better understand harmonized maps.

## References

Bucher B, Brasebin M, Buard E, Grosso E, Mustière S and Perret J, 2012, GeOxygene: Built on top of the expertness of the french NMA to host and share advanced GI science research results. In: Bocher E, Neteler M (eds) Geospatial Free and Open Source Software in the 21st Century. LNG&C, Springer, Berlin, 21–33.

Harrie L, Sarjakoski T, 2002, Simultaneous graphic generalization of vector data sets. Geoinformatica 6(3):233–261.

Sester M, Jokar Arsanjani J, Klammer R, Burghardt D and Haunert JH, 2014, Integrating and generalising volunteered geographic information. In: Burghardt D, Duchêne C and Mackaness W (eds), Abstracting Geographic Information in a Data Rich World, LNG&C. Springer, Berlin, 119–155.

Touya G, 2012, What is the level of detail of OpenStreetMap? In: Workshop on Role of Volun-teered Geographic Information in Advancing Science: Quality and Credibility, Columbus, USA.

Touya G, Brando C, 2013, Detecting Level-of-Detail inconsistencies in volunteered geographic information data sets. Cartographica 48(2):134–143.

Touya G, Coupé A, Le Jollec J, Dorie O, Fuchs F, Conflation optimized by least squares to maintain geographic shapes. ISPRS International Journal of Geo-Information 2(3):621–644.

# A Humanities GIS Ontology: *Tweetflickertubing* James Joyce's *Ulysses* (1922)

Charles Travis[1]

[1]Trinity Long Room Hub, Trinity College, Dublin 2, Ireland
Email: ctravis@tcd.ie

## 1. Introduction

This Humanities GIS (HGSI) model cross-pollinates literary and social media practices to engage in a participatory, performative and augmented reality survey of the relations between James Joyce's novel Ulysses (1922) and digital eco-system productions of dialogical and social space. (Goodchild 2009; Sieber, Wellen, and Yuan 2011; Priem, 2011; Young and Gilmore 2013; Graham and Zook, 2013; Lin, 2013). Joyce famously boasted that the aim of Ulysses (which kaleidoscopically relates the urban journeys of student Stephen Dadelus and advertising salesman Leopold Bloom on June 16th 1904) was "to give a picture of Dublin so complete that if the city one day suddenly disappeared from the earth it could be reconstructed out of my book" (Budgen, 1972, 69). The model integrates a Ulysses schema outline with live geo-spatially enabled Twitter, Flickr and YouTube posts to map the language operating in a Bloomsday generated digital eco-system to recreate the discourse of a virtual Joycean Dublin during the annual celebration of the novel. Consequently the 'Joycean' neologism tweetflickertubing was coined to describe the ontological shift indicated by the HGIS model's methodology.

### 1.1 Creating the model

In 1920 Joyce drafted a schema outlining Ulysses' eighteen Homeric episodes for the Italian critic Carlo Linati to whom he wrote: "in view of the enormous bulk and the more than enormous complexity of my damned monster-novel it would be better to send . . . a sort of summary-key-skeleton schema" (Ellman, 1974, 187). The schema's grid designates episode title, time, color, people, science/art, meaning, technic, organ, and symbols. To create the model an Excel spreadsheet (Figure 1) was populated with this data and geo-coded according to GPS designated decimal degree locations of the 18 Homeric episodes sites in Dublin. These sites were identified through the 1904 *Thom's Map of Dublin*, Ian Gunn and Clive Hart's, *James Joyce's Dublin: A Topographical Guide* (2004), GIS 'ground-truthing' methods and checked against 'what's here' function of the Google Maps App.

Site centroids were approximately identified to concrete locations described in Ulysses because the various characters' movements and locations within the novel (such as the Wandering Rocks episode) occur simultaneously and at multiple sites within and beyond the geographical and temporal boundaries distinguishing each episode in the schema. The Excel database was imported into Google Fusion, and visualized through its Google Maps function. The database was also converted to a CSV file and imported into the ArcGIS Online platform and integrated with a live social media map layer.

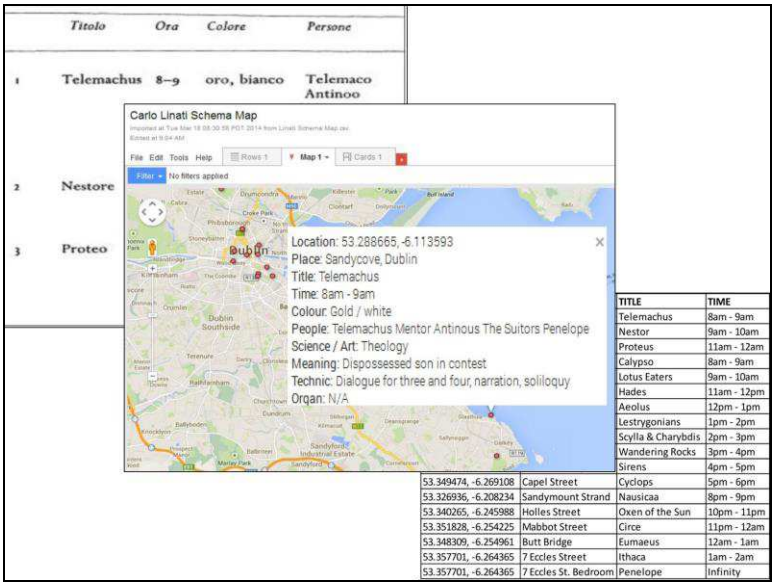



Figure 1: Fragments of the Carlo Linati Schema, Google Fusion Map and Excel/CVS Geo Database.

## 2. Surveying Bloomsday Digital Ecosystem

Surveys were taken on June 16th, 2014 at hourly intervals, based upon the chronology from the Linati Schema, and were divided into two categories –local (Dublin) and global, weighed by the indices of site and time respectively. Several keywords such as "James Joyce," "Ulysses", and "Bloomsday" as well as character and episode names from the novel were tested in the model's Twitter, Flickr and YouTube search engines. "Bloomsday" received the highest number of hits and became the main survey keyword.

The local survey focused on activity around Homeric episode sites in Dublin, and found that Davy Byrne's Pub on Duke Street and Joyce's Martello Tower associated with the Lestrygonians and Telemachus episodes attracted the most posts. Tweet postings did not however, did not correspond with the chronology of *Ulysses'* narrative outlined in the schema, illustrating that social media activity aggregated around site location, rather than novelistic time. A Tweeted image (Figure 2) captures throngs of people in funny hats assembled around Byrne's pub and it seems Joyce's identification of the 'Oesophagus' as the body organ symbol for this episode was indeed apt. The National Library site was originally geo-coded as the centroid of the Lestrygonians episode, however survey results suggests that perhaps because of the social gravity indicated by number of social media posts, the centroid should be re-located, to the site of the pub, illustrating the iterative process integral to Neogeographical mapping practices. In the case of the global survey, Tweets blossomed across Europe, the Middle East, Asia, Australia-Pacific, North and Latin America during the entire chronology of the Linati Schema. 'Orphan' Tweets (corresponding outside of the hourly periods not included in the schema) were placed in either in preceding episode time slots, or in the Penelope episode- whose time indices encompasses 'Infinity' (see Bloomsday Tweet Schema in poster).


Figure 2: Live social media map integrated with the Linati Schema geo-database.

Surveys taken before, on and following Bloomsday 2014 illustrate that Flickr and YouTube postings with time lags, and reflecting activity over the course of a year exhibited the most aggregated social media activity. However, over the course of the June, 16th 2014 survey, it became apparent that dominant social media ecosystem activity on Bloomsday occurred in Twitter. The following verbal snapshot reflects a parsing of language activity (see Bloomsday Tweet Schema in poster) articulated in this digital eco-system: "People in Dublin are wearing funny hats because it is Bloomsday and elderly ladies are getting rowdy in Davy Byrne's Pub; a wedding anniversary is observed in Glasnevin, North Dublin, while a Spanish tweeter celebrates with a Domino's Pizza and the latest X-Men film. Individuals in Dublin, Paris and Washington D.C. resolve to again attempt to read Ulysses, and a tweeter in Uruguay mentions Bloomsday to her Irish boyfriend, who asks if the day has anything to do with flowers. A few literary minded types post Joycean lines from the novel, while two individuals from Dublin get suited up in Edwardian clothes to face the day; one tweeter reflecting on the day after the night before, asks if Bloomsday was a joke brought up in a drinking session. A tweeter from Mexico City advertises online translations of Joyce's 'lascivious' letters to his wife Nora Barnacle. The celebration of Ulysses converges with perhaps a larger global event to provoke a Dublin tweeter to state that there is 'Nothing like a combo of World Cup and bloomsday to hear people who don't like either Joyce or football talk about both.' One wry observation from the Bronx asks if 'Bloomsday is Paddy's Day for posh people?' And two more tweets from the USA proclaim 'I'm pretty sure Joyce would love hashtags,' and 'To paraphrase Laurie Anderson: Ulysses? Never read it.'"

A corollary can be made between Joyce's writing technique in Ulysses and the use of language in this Twitter based digital ecosystem. Joyce's stream of consciousness technique mimicked the various ways in which the human mind 'speaks' to itself, through complex fluid patterns, random interruptions, incomplete thoughts, half words and tangents (Norris and Flint, 2000, 126). Tweets, limited to a certain numbers of characters reflect Joyce's technique by conveying both focused and random thoughts, and illuminate Graham and Zook's (2013, 78) contention that the "digital dimensions of places are fractured along a

number of axes such as location, language, and social networks with correspondingly splintered representations customized to individuals' unique sets of abilities and backgrounds." The HGIS model illuminates how literary and historical tropes can aid in contextualizing and mapping social media activity through the creation interpretive schemas to study interactions between language, behaviour, time and place.

## 3. HGIS implications

Re-conceptualizing J.B. Harley's observation that "text" is a better metaphor for maps than the mirror of nature through the lens of the digital humanities, it can be seen that HGIS generated "maps are a cultural text. By accepting their textuality we are able to embrace a number of different interpretative possibilities" (1989, 4). The HGIS model enabled digital inter-textual relationships between Ulysses, the Linati Schema and the dialogical and social space reflected in the Bloomsday social media eco-system. Subsequently, digital humanities methodologies of deformance and ergodicity were applied as interpretative techniques. By translating one ontological form of discourse to another, deformance applies scientia to poeisis and seeks to explain unitary and unique phenomena (such as the language activity in the Bloomsday social media ecosystem) rather than establish a set of general rules or laws (McGann and Samuels, 1999). Ergodicity involves an interactive type of labor between the GIS practitioner/author/coder, reader/viewer and mapping subject to create the potential multiple narrative paths composing a digital text. Ergodic applications of GIS create non-linear narratives which converge and disrupt both quotidian and epochal chronologies of time and space. As a literary tool this HGIS model synchronizes the resulting layers of images, words, and vectors into contrapuntal, multi-dimensional digital narratives, providing the means "reconnect the representational spaces of literary texts not only to material spaces they depict, but also reverse the moment" (Staley 2007; Thacker, 2005, 63). Lastly, HGIS, Digital, Spatial, and Geohumanities modelling techniques, integrated with radical statistics holds the potential to engage Qualitative & Critical GIS/ GIScience studies with reflexive epistemologies to address the situatedness and positionality of Big Data as it relates to sustainability initiatives, smart city planning, transport and public health, disaster preparedness and social and regional conflict (Cope and Elwood, 2009; Aitken and Craine, 2009; Bodenhamer, et. al., 2010; Dear, et. al. 2011; Meir 2011; Elwood, et. al., 2012; Kwan and Schwanen, 2012; Leventala, 2012; Kitchin, 2014; Travis 2014).

## References

Aitken, S. and Craine, J. 2009. Into the Image and Beyond: Affective Visual Geographies and GIScience. In: S. A. Elwood and M. Cope, (eds), Qualitative GIS: A Mixed-Methods Approach. Sage, London.

Bodenhamer, D., Corrigan, J. and Harris, T.M., 2010, The Spatial Humanities, Indiana University Press, Bloomington & Indianapolis.

Budgen, F., 1972, James Joyce and the Making *of 'Ulysses'* Oxford University Press, Oxford.

Bulson E, 2001-2002, Joyce's Geodesy. Journal of Modern Literature 25(2): 80-96

Cope, M. and Elwood, S.A., 2009. Qualitative GIS: A Mixed Methods Approach. Sage, London.

Dear, M., Ketchum, J. Luria, S. and Richardson, D. 2011, Geohumanities: Art, history, text at the edge of place, Routledge, N.Y.

Ellman, R. 1974.Ulysses on the Liffey. Faber & Faber, London.

Graham, M., and Zook, M., 2013, Augmented realities and uneven geographies: exploring the geo-linguistic contours of the web, Environment and Planning A 45: 77-99

Goodchild, M. 2009. NeoGeography and the nature of geographic expertise. Journal of Location Based Service. 3 (2): 82-96

Gunn, I. and Hart, C. 2004. *James Joyce's Dublin: A To*pographical Guide. Thames and Hudson, London.

Harley, J.B. 1989. Deconstructing the Map. Cartographica. 26 (2):1-20.
Kitchin, R. 2014. Big Data, new epistemologies and paradigm shifts, Big Data and Society (1):1-12.
Kwan, M. P. and Schwanen, T. 2012. Critical Space-Time Geographies. In: Tim Schwanen and Mei-Po Kwan. (eds) Environment and Planning A, 44 (9): 2043-2048
Leventala, S. 2012. A New Geospatial Services Framework: How Disaster Preparedness Efforts Should Integrate Neogeography, Journal of Map & Geography Libraries: Advances in Geospatial Information, Collections & Archives. 8(2): 134-162
Lin,W. 2013. Situating performative neogeography: tracing, mapping, and performing "Everyone's East Lake". Environment and Planning A. (45): 37-54
McGann, J. and Samuels, L. 1999. Deformance and Interpretation. New Literary History. 30 (1): 25-56
Meir, P. 2011. New information technologies and their impact on the humanitarian sector. International Review of the Red Cross. 93 (884): 1239-1263
Norris, D. and Flint, C. 2000. Introducing Joyce. Icon Books, Cambridge.
Priem, J. 2011. As Scholars Undertake a Great Migration to Online Publishing, Altmetrics Stands to Provide an Academic Measurement of Twitter and Other Online Activity. Impact of Social Science. [http://blogs.lse.ac.uk/impactofsocialsciences/2011/11/21/altmetrics-twitter/] Accessed May 2013.
Sieber, R.E.; Wellen, C.C.,Yuan J. 2011. Spatial cyberinfrastructures, ontologies, and the humanities. Proceedings of the National Academy of Sciences of the United States of America. (108) 14: 5504-5509.
Staley, D. J. 2007. Finding Narratives of Time and Space. In: D.S. Sinton & J.J Lund, (eds.) Understanding Place: GIS and Mapping Across the Curriculum, Esri Press, Redlands: 35-48
Thacker. A. 2005. The Idea of a Critical Literary Geography. New Formations. 57 (6): 56-73.
Travis C. 2014. Transcending the Cube: Translating GIScience Time and Space Perspectives in a Humanities GIS, Special Issue on Space-Time Research in GIScience, International Journal of Geographical Information Science. 28(5): 1149-1164
Young, J.C. & Gilmore, M.P. 2013. The Spatial Politics of Affect and Emotion in Participatory GIS. Annals of the AAG.103 (4): 808-823.

.

# Using GIS and Geo-targeted Social Media (Twitter) to Track Illicit Drug Use Trends in Space and Over Time

Ming-Hsiang Tsou[1], Susan I. Woodruff[2], Brian Spitzberg[3], Mark Reed[2], Meghan Moran[3], Mark Gawron[4] , Christopher. Allen[1], Jiue-An Yang[1]

[1]San Diego State University, Department of Geography,
5500 Campanile Drive, San Diego, CA 92182-4493
Emails: mtsou@mail.sdsu.edu, ccchris.allen@gmail.com, jyang@mail.sdsu.edu

[2]San Diego State University, School of Social Work,
5500 Campanile Drive, San Diego, CA 92182-4493
Email: swoodruff@mail.sdsu.edu, mreed@mail.sdsu.edu

[3]San Diego State University, School of Communication,
5500 Campanile Drive, San Diego, CA 92182-4493
Email: spitz@mail.sdsu.edu, mmoran@mail.sdsu.edu

[4]San Diego State University, Department of Linguistics,
5500 Campanile Drive, San Diego, CA 92182-4493
Email: gawron@mail.sdsu.edu

## 1. Introduction

The spread of social media in society is increasing rapidly, and surveillance based on such sources offers the potential to monitor illicit drug use trends and even emerging drug use. Based on the 2013 Pew Research Center Survey (http://www.pewinternet.org/), over 90% of young Internet users (age 18-29) are using social networking sites. Twitter has been generally recognized as the social network and social media for younger generation. Older teens and young adults are the heaviest Twitter users, which approximately 74% of Twitter users are in the age range of 15-25 (www.beevolve.com). People within the highest Twitter age range also report the highest rates of illicit drug use, according to national surveys (NSDUH, 2013). About 60% of this age range report lifetime use, 37% report past year use, and 25% report past-month use of illicit drugs. The use of Twitter data upon which to base surveillance is practical in that it captures information from the age group most at risk. Despite the illicit nature of some drug activities, research shows that people nevertheless communicate online about such activities (Mackey & Liang, 2013), and this communication is available for near real-time surveillance (Spitzberg, in Press).

This research adopted GIS methods, natural language processing, machine learning algorithms, and advanced geo-targeted social media application programming interfaces (APIs) to track drug use trends (including new and emerging drugs) in space and over time. A prototype system, called Spatial, Temporal, and Regional Observation Network Generator for Drug Abuse Trend Analysis (STRONG-DATA), is under development with an iterative approach to collect, filter, and analyze user-generated content from Twitter.

## 2. Adopting Knowledge Discovery in Cyberspace (KDC) framework for Tracking Illicit Drug Use in U.S.

Our research team has developed a comprehensive data process/analysis research framework for studying social media, called Knowledge Discovery in Cyberspace (KDC) (Tsou and Leitner, 2013). The KDC framework is used to collect and analyze social media messages related to illicit drug use, which are distributed in different places, different times, and with different networks. We use GIS and geo-tagged social media APIs to collect and analyze the

dynamic relationships between time, place, and messages and build a triangular knowledge base for understanding and analyzing illicit drug use trends (Figure 1).

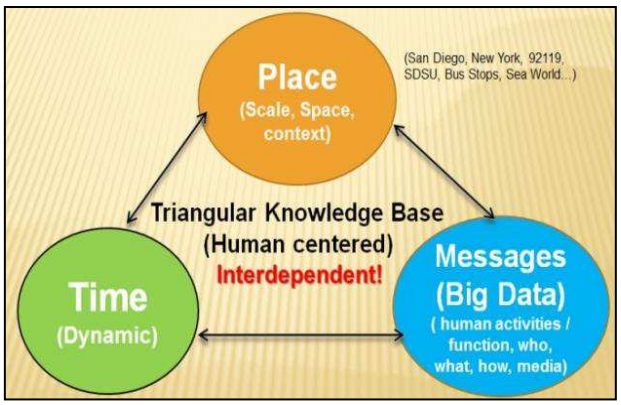



Figure 1: The Knowledge Discovery in Cyberspace (KDC) Framework (Tsou and Leitner, 2013).

Our geo-targeted search APIs collects two types of spatial information from tweets: 1. geo-tagged locations provided by GPS-enabled devices, 2. self-reported locations specified in user profiles. Geo-tagged locations are latitude and longitude pairs created by mobile devices with built-in GPS receivers or by other geo-location features. The self-reported location is specified by users and can be changed by users at any time. Each social media message is collected with detailed attributes including user_id, text content or media contents, created_time, and spatial locations, which includes geo-tagged coordinates or self-reported place names. Python, R, and JavaScript are used to create social media analytics tools and automatic filter procedures.

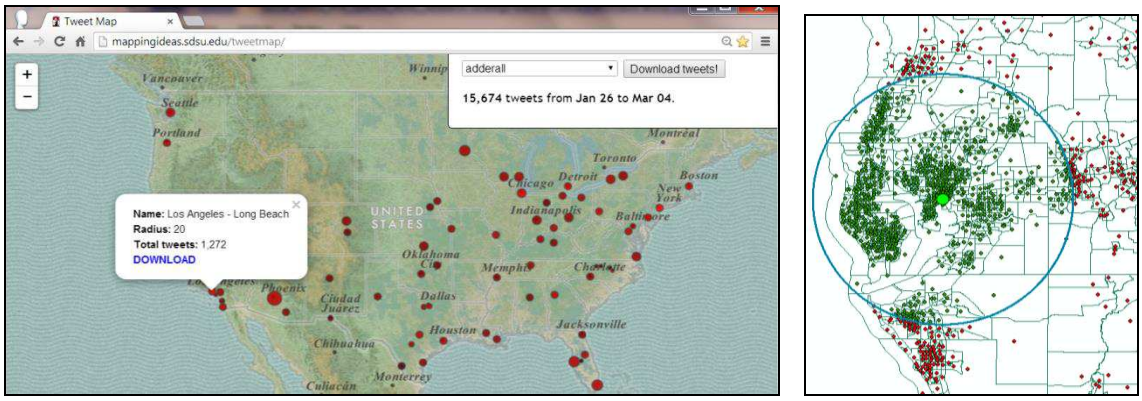



Figure 2. Using the geo-targeted APIs (left) and GIS methods (right) to collect tweets from U.S. cities.

Figure 2 illustrates the collection of the keyword "Adderall" from the 100 largest U.S. cities in the geo-targeted API tool. This tool can continue to search relevant tweets daily based on a list of keywords and a list of city names with user-defined radius. The collected tweets in each city are standardized by its population within the radius circle using 2010 census data and GIS tools (Figure 2). In our pilot study, we collected tweets using three keywords, "Adderall", "Kush", and "Heroin" (non-case sensitive searches) from 10 selected U.S. cities within 20 miles radius from January 26, 2014 to March 11, 2014 (37 days total). The design of STRONG-DATA integrates natural language processing and machine learning algorithms to filter noise from the original tweet data collection. A preliminary filtering procedure of removing retweets, URL-tweets, and tweets with similar user names is used in the first phase of STRONG-DATA. The geo-targeted APIs collected 8698 tweets for the "Adderall" keyword and the preliminary filtering procedure reduced the dataset from 8698 tweets to 4842 valid tweets. The other keywords ("Kush" and "Heroin") had similar results. The second phase is under development and uses machine learning algorithms that could further improve the filtering results. Figure 3 illustrates the word cloud Image (created by R) by combining all 4842 tweet texts for "Adderall", 16828 tweet texts for "Kush", and 3714 tweet texts for "Heroin". The word cloud Images provide some context explanation of tweets

related to different drug use (Adderall, Kush, and Heroin).

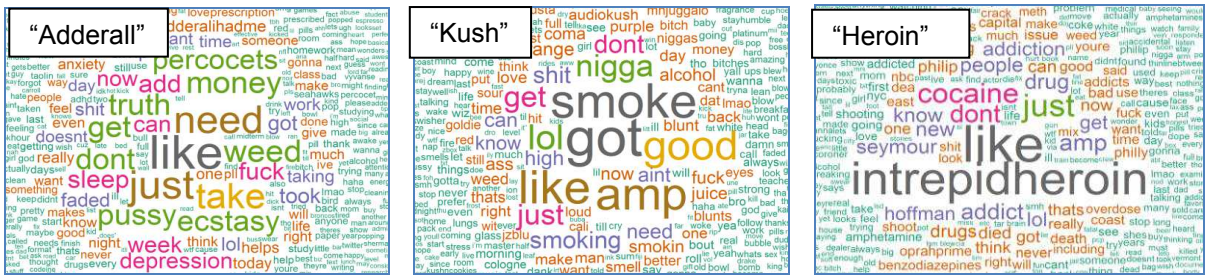



Figure 3. The social media messages (word cloud images) from 4,842 "Adderall" tweets (left), 16,828 "Kush" tweets (middle), and 3,714 "Heroin" tweets (right) from 10 U.S. cities combined. (1/26/2014 – 3/11/2014).

By adapting the KDC framework, we analyzed the spatial distribution pattern of these tweets in different places (cities) and different times (weeks). Figure 4 illustrates the "Adderall" tweeting rates in 10 U.S. cities (left) and the "Kush" tweeting rates in 10 U.S. cities (right). We found that Boston had the highest tweeting rates of "Adderall" compared to other cities, but only ranked 5th in "Kush" tweeting rates. Atlanta had the highest "Kush" tweeting rate, but only ranked 6th in "Adderall" tweeting rate. Our conclusion was that different cities have their own patterns/signatures of tweeting about illicit drug use. In other words, the term "Kush" may be a regionally specific way of referring to marijuana or individuals in Atlanta may in fact be using more marijuana. Traditional methods of social media data capture and analysis are not sophisticated enough to differentiate between regional variance in terminology and use.

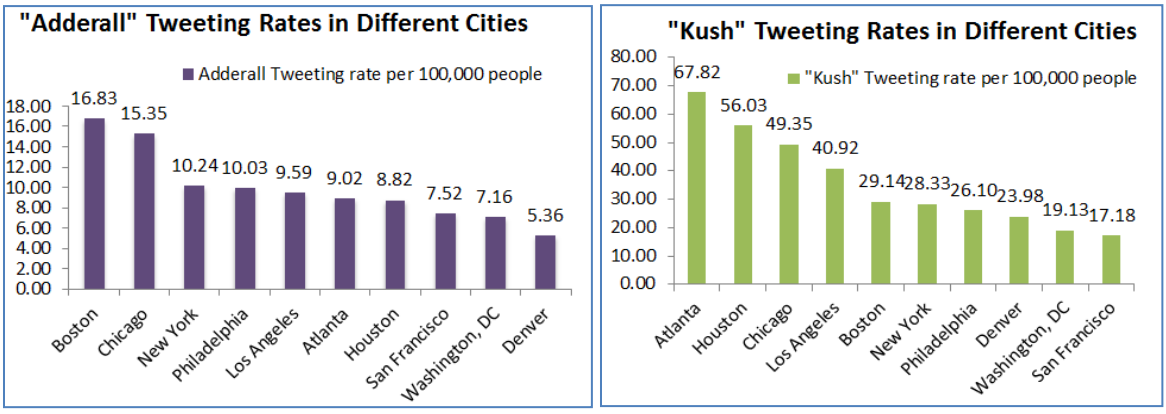



Figure 4. "Adderall" and "Kush" tweeting rates by city (1/26/2014 – 3/11/2014).

From a spatiotemporal perspective, we analyzed the weekly trend of "Adderall" tweets in different cities and in weekday/weekend. Figure 5 (left) illustrates the temporal pattern of "Adderall" tweets in each week (from 1/26/2014 to 3/11/2014) from New York City and Los Angeles. The peak week of tweeting in New York City was 2/16-2/22. The peak week in Los Angeles was 3/2-3/8. Figure 5 (right) illustrates another type of temporal patterns: both New York City and Los Angeles had the highest peak of "Adderall" tweeting rates on Tuesday and the lowest tweeting rates on Saturday (possibly due to use of the drug for academic or work schedules). The unique ability of the STRONG-DATA system to track temporal trends across geographies, and to track geographical trends across time has important implications for intervention. Healthcare providers and public health officials may be able to use these data to allocate resources in an efficient and effective way.

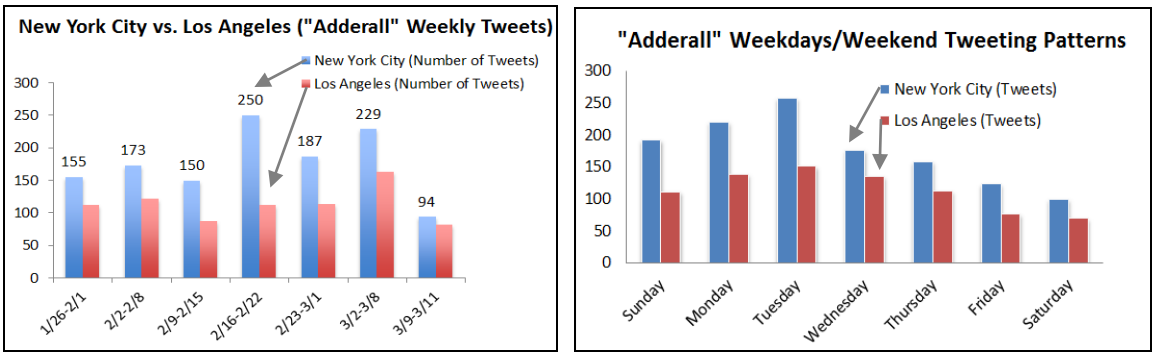



Figure 5. Temporal patterns of "Adderall" tweeting rates in NYC and Los Angeles

## 3. Confidentiality and Ethical Issues

One important aspect of this research is the ethical issue about personally identifiable (directly or indirectly) health-related and/or substance use/abuse information. To protect the privacy of Twitter users, a number of strict data protection protocols will be followed in our study. Our processed Twitter data will reside on a Level-1 secure university data server with state-of-the-art multilevel security. We will leverage multiple privacy protection methods in privacy graph computation (Gao et al. 2011), text-based privacy, and locational privacy.

## Acknowledgements

This study is partially supported by the National Science Foundation under Grant No. 1028177, project titled “Mapping Cyberspace to Realspace”. The authors thank Anoshe Aslam for her data processing efforts.

## References

Gao, J., Xu Yu, J., Jin, R., Zhou, J., Wang, T., & Yang, D. (2011) Neighborhood-Privacy Protected Shortest Distance Computing in Cloud. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 409-420. New York, NY: Association Computing Machinery, 2011.

Mackey, T. K., & Liang, B. A. (2013). Global reach of direct-to-consumer advertising using social media for illicit online drug sales. Journal of Medical Internet Research, 15(5), e105, 1-14.

National Survey on Drug Use and Health (NSDUH). (2013). Results from the 2012 National Survey on Drug Use and Health: Summary of National Findings. NSDUH Series H-46, HHS Publication No. (SMA) 13-4795. Rockville, MD: Substance Abuse and Mental Health Services.

Spitzberg, B.H. Toward a model of meme diffusion ($M^3D$). Communication Theory, in press.

Tsou, M-H., & Leitner, M. (2013). Visualization of social media: seeing a mirage or a message? Cartography and Geographic Information Science, 40(2).

Tsou, M-H., Yang, J-A., Lusher, D., Han, S. H., Spitzberg, B. H., Gawron, J. M., Gupta, D., An, L. (2013). Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): A case study in the 2012 U.S. Presidential election. Cartography and Geographic Information Science, 40(4), 337-348.

# Cluster Detection in Networks by controlling Shape Flexibility Level

M. Tsukahara[1], R. Inoue[1]

[1]Graduate School of Information Sciences, Tohoku University, Aramaki Aoba 6-6-06, Aoba, Sendai, Miyagi, 980-8579 Japan
Email: {tsukahara; rinoue}@plan.civil.tohoku.ac.jp

## 1. Introduction

Cluster detection of point events is a regional analysis method. Several approaches for detecting clusters have been proposed; most of them are based on the spatial scan statistic proposed by Kulldorff and Nagarwalla (1995). The spatial scan statistic is a maximum likelihood ratio test statistic. It configures candidates of a cluster, computes the likelihood ratio as an index of degrees of the point accumulation for each candidate, and detects the candidate with the maximum likelihood ratio value as a cluster. A variety of methods based on spatial scan statistics are identified by the settings of the cluster candidate shapes, which differ from fixed to flexible. These circular (Kulldorff 1997), elliptic (Duczmal et al. 2006), and flexible shapes are formed by the combination of neighbouring regions (e.g. Duczmal and Assunção 2004, Duczmal et al. 2007). Furthermore, Duczmal et al. (2006) proposed controlling the shape flexibility of clusters that consist of the set of neighbouring regions.

Recently, Shiode (2011) proposed a cluster detection method for networks that is an expansion of spatial scan statistics in network analysis. Because this method is capable of describing the micro-space variation of locations of point events at the street level, it is suitable for analysis of detailed location information. However, it lacks flexibility in detecting cluster shapes; it can only detect a 'circle-like' cluster in a network, which is defined by the network distance from a specific point. Because not all clusters in a network are caused by isotropic propagation processes, this method offers limited application in detecting clusters in networks.

This paper proposes a new method to detect flexibly shaped clusters as sets of connected links in networks while controlling the flexibility level.

## 2. Controlling cluster shape flexibility in region-based analysis: previous research

Duczmal et al. (2006) introduced compactness value to control the shape flexibility of clusters when searching for clusters as sets of neighbouring regions. Let Z denote a cluster candidate set by the combination of neighbouring regions, A(Z) denote the area of Z, and H(Z) denote the perimeter of a convex hull of Z. The compactness value K(Z) is defined by the ratio of A(Z) and the area of a circle with perimeter H(Z).

$$K(Z) = A(Z) \Big/ \pi\left(\frac{H(Z)}{2\pi}\right)^2 \tag{1}$$

K(Z) is the dimensionless shape index of Z. Its value ranges between zero and one; it approaches one when Z has a circular compact shape.

Duczmal et al. (2006) used $LR(Z)^{K(Z)^a}$ as the evaluation function of cluster candidates instead of the likelihood ratio $LR(Z)$ to control the flexibility of cluster shapes. a is a user-specified scaling value; no constraints are imposed on the cluster shape when a is zero, and the constraints on the shape become strict as a increases. Analysts can control the cluster shape flexibility by setting the value of a.

## 3. Cluster detection in networks by controlling shape flexibility

In region-based analysis, the compactness value is defined by the ratio of areas. The 'length' of links in network analysis is equivalent to the 'area' of regions in region-based analysis; thus, this study defines the compactness value in networks, $K_n(Z)$, by the length of the cluster candidate link set Z.

Before discussing compactness value in networks, it is necessary to address the compact shape in networks. This study defines the compact shape in networks as the link set for which all distances from the central node to each endpoint vertex are equal. This property is referred to as the property of a circle in the Euclidean space in which 'all distances from the centre to the circumference are equal'. Figure 1 shows an example of a compact shape: the red circle depicts the central node, blue lines represent edges, and the network distances from the central node to the endpoint vertices of the link set are equal.

This study evaluates the compactness of Z by comparing it to its corresponding compact shape. Let C(Z) denote the compact shape link set corresponding to Z; it has the same total length and the same central node as Z. This study defines the central node of a network as the node whose network distance to the farthest node is the shortest. It is similar to the centre of the smallest enclosing circle in the Euclidean space. After the central node of Z is identified, C(Z) is generated by selecting the connected links in short-distance order from the central node until the total length exceeds that of Z.

At this point, let D(X) denote the network distance to the farthest node from the central node in a certain link set X. Then, the compactness value in networks $K_n(Z)$ is given by

$$K_n(Z) = \frac{D\{C(Z)\}}{D(Z)}. \tag{2}$$

Figure 2 illustrates the relationship between $K_n(Z)$ and the compactness of the link set. $K_n(Z)$ approaches one when Z is compact and zero when Z has a complex shape.

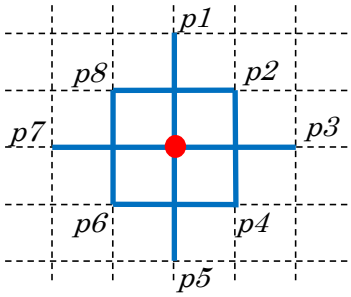



Figure 1: Example of a compact link set.

(a): Compact-shaped network ($K_n(Z) = 0.91$). (b): Complex-shaped network ($K_n(Z) = 0.30$).

Figure 2: Relationship between $K_n(Z)$ and network shape complexity.

In this study, we modify the genetic algorithm scan of the region-based analysis of Duczmal et al. (2007) to fit network analysis. By limiting the shape of cluster candidates in the search process using the compactness value in the network, the proposed method controls the shape flexibility of detected clusters.

## 4. Application

The proposed method was applied to the geographic distribution of restaurants in Aoba ward, Sendai city. The road network data is from the Open Street Map; it consists of 11,368 links with a total length of 1,358 km. The restaurant location data, obtained from the "Telepoint Pack Database" telephone directory of February 2011, includes the geographic coordinates of 501 restaurants. Table 1 and Figure 3 show the cluster detection results with shape constraints. Grey lines represent road networks, green dots depict the locations of restaurants, and red lines indicate the detected clusters. Figure 3(a) illustrates the result obtained with no shape constraint; the detected cluster covers a broad area and has a complex shape. Figures 3(b)-(d) illustrate the results obtained with shape constraints; the detected clusters become small and compact as the shape constraint becomes strict. These results demonstrate that the proposed method can adjust the shape flexibility of clusters by manipulating a constraint setting based on the compactness value in networks.

Figure 4 and Table 2 show the result of detecting the primary cluster and two additional non-overlapping clusters with minimum $K_n = 0.8$. Red lines indicate the primary cluster, blue lines indicate the secondary cluster, and green lines indicate the tertiary cluster. The set of these clusters is equivalent to the cluster in Figure 3(a). The cluster detection method found three distinct compact clusters that together formed one large cluster when run using strict shape constraints.

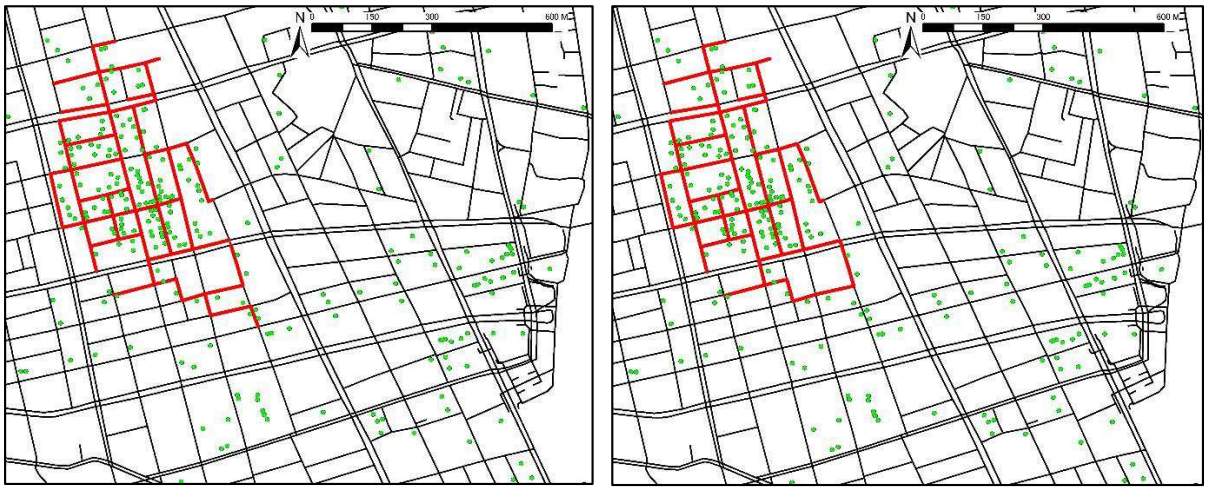
Table 1: Cluster detection results with shape constraints.

| Case | Constraints for minimum $K_n$ | Likelihood ratio | $K_n$ of detected clusters |
|---|---|---|---|
| 1 | none | 1,445 | 0.35 |
| 2 | 0.6 | 1,116 | 0.60 |
| 3 | 0.7 | 1,042 | 0.71 |
| 4 | 0.8 | 1,030 | 0.80 |

Table 2: The primary cluster and two non-overlapping clusters with minimum $K_n = 0.8$.

| Cluster | Likelihood ratio |
|---|---|
| Primary | 1,030 |
| Secondary | 264 |
| Tertiary | 124 |

(a) Case 1: no shape constraint. (b) Case 2: minimum $K_n = 0.6$.



(c) Case 3: minimum $K_n = 0.7$. (d) Case 4: minimum $K_n = 0.8$.

Figure 3: Primary cluster detection results with shape constraints.

Figure 4: The primary cluster and two non-overlapping clusters with minimum $K_n = 0.8$.

## 5. Conclusion

In this paper, we proposed the use of the compactness value in networks to control the shape flexibility of detected clusters. We applied this approach to a cluster detection method. The application revealed that the proposed method succeeds in controlling the shape flexibility of detected clusters in networks.

## Acknowledgements

The "Telepoint Pack Database" provided by Zenrin Co., Ltd. was used as part of the CSIS Joint Research (No. 456) of the Centre for Spatial Information Science, the University of Tokyo.

## References

Duczmal L and Assunção R, 2004, A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistics and Data Analysis, 45:269–286.

Duczmal L, Kulldorff M and Hung L, 2006, Evaluation of spatial scan statistic for irregularly shaped clusters. Journal of Computational and Graphical Statistics, 15(2):428–442.

Duczmal L, Cançado A, Takahashi R, and Bessegato L, 2007, A genetic algorithm for irregularly shaped spatial scan statistics. Computational Statistics and Data Analysis, 52:43–52.

Kulldorff M and Nagarwalla N, 1995, Spatial disease clusters: Detection and inference. Statistics in Medicine, 15:707–715.

Kulldorff M, 1997, A spatial scan statistic. Communications in Statistics-Theory and Methods, 26:1481–1496.

Shiode S, 2011, Street-level spatial scan statistic and STAC for analysing street crime concentrations. Transactions in GIS, 15(3):365–383.

# A time series analysis of land cover change: random forest models of annual changes in urban land cover

Narumasa Tsutsumida[1,2], Alexis J. Comber [2], Kirsten Barrett[2], Izuru Saizen[1]

[1]Graduate School of Global Environmental Studies, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto,Japan
Email: {naru; saizen}@kais.kyoto-u.ac.jp

[2]Department of Geography, University of Leicester, Leicester, UK
Email: {ajc36; kb308}@leicester.ac.uk

## 1. Introduction

Huge volumes of spatio-temporal data have been collected by continuous earth observation satellite sensors such as MODerate resolution Imaging Spectroradiometer (MODIS). Such data are useful for temporal analyses that require consistent sampling intervals. For example, MODIS MOD13Q1 product provides Enhanced Vegetation Index (EVI) datasets. This product is a 16-day composite product at 250-meter spatial resolution, composed of the high quality pixels in terms of cloud-cover, view angle, and residual atmospheric contamination over the period. GI systems are able to manage spatial data, examine spatial relationships and to understand spatial processes, but they are lacking in their treatment of both spatial and temporal dimensions. However as geographic phenomena constantly change over time, knowledge extracted from spatio-temporal data analyses can help us gain better insight into spatial processes (Yao 2003).

One of the best examples of spatio-temporal changes on land surface is urban expansion. Although much research has monitored urban expansion using remotely sensed data, these analyses have typically used data collected at irregular or temporally coarse intervals, for example using Landsat and SPOT data. However, there remains an unmet need for observations at high temporal frequency (Sexton et al. 2013). This research develops a method that produced annual land cover maps and identifies annual land cover changes using remotely sensed data collected at regular intervals. We focuses on the annual time-series pattern of EVI signal. By considering the differences among such patterns, it is theoretically possible to determine changes in land cover related to urban expansion. It uses MODIS data to evaluate land cover change in a study area around Jakarta and its suburban area, Jabodetabek, in Indonesia. This region covers a total area of 6,700 $km^2$ and represents one of the largest mega-cities in south-east Asia and is expanding continuously according to economic growth and population statistics.

## 2. Data

This analysis used 299 images of EVI covering the period 2001–2013, provided by the MOD13Q1 product (Version 5). The study area consists of 63,477 pixels in a EVI imagery, so totally 18,979,623 pixels (299 images $\times$ 63,477 pixels) were prepared. Some1,000 randomly selected areas of the same resolution as the MOD13Q1 pixel were sampled to generate ground truth data. The land cover proportions at these areas were visually interpreted using available very high spatial resolution (VHR) imagery in Google Earth during the study period. This resulted in 4,642 reference points at several time periods.

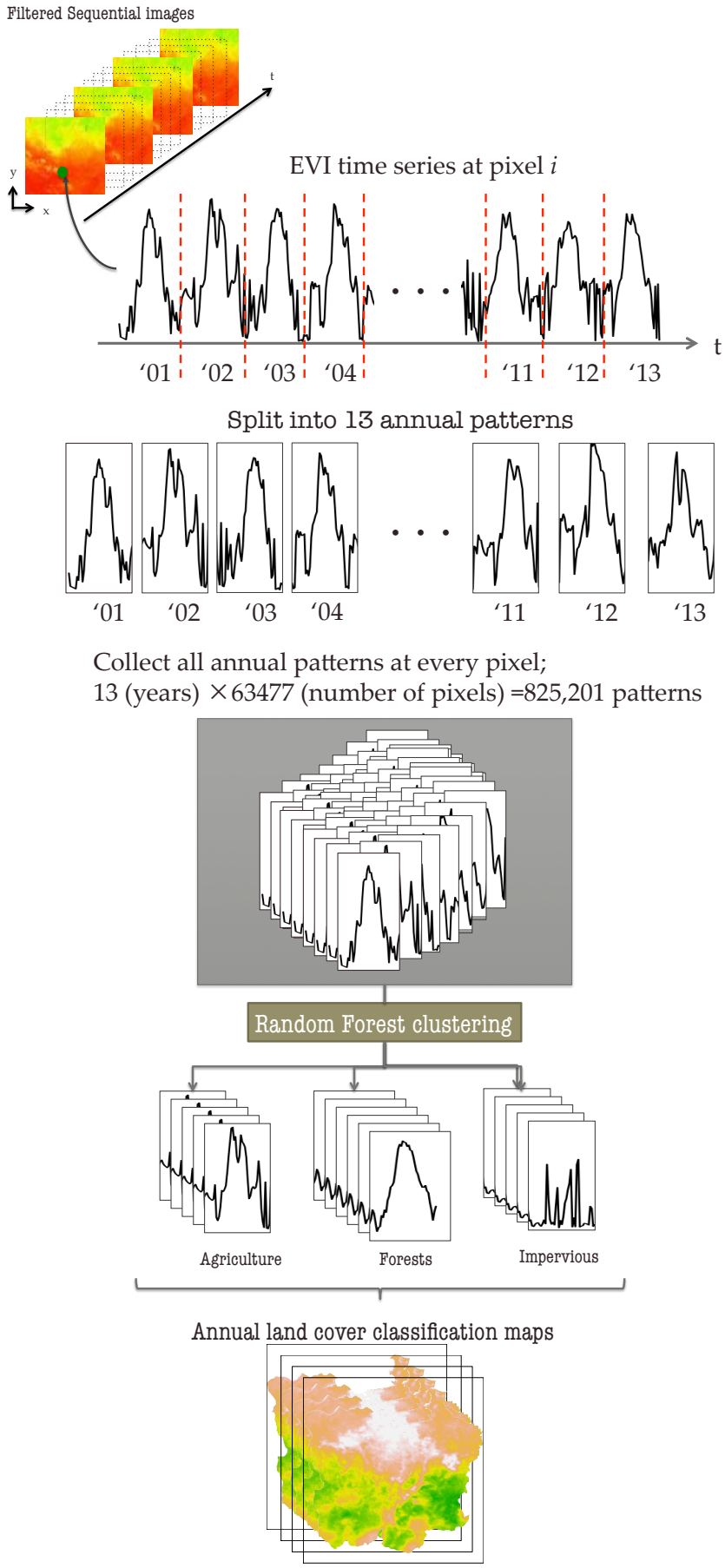



Figure 1: Procedure of analysis.

## 3. Methods

All reliable data shown as good or marginal flag at MOD13Q1 reliability layer were extracted from original EVI time series. Then, NA values were interpolated using double logistic function in TIMESAT (Jönsson and Eklundh 2004). The annual time-series patterns were extracted (13 years × 63,477 pixels) providing 825,201 patterns available for analysis. The number of legend categories was set to 3 because land cover in the study area is mainly composed of impervious surface, agriculture, and forest. Reference data with greater than 70% coverage by any single land cover were only selected (2,048 samples), of which 50% were used to training the classifier. Using them, a random forest classifier was applied to all patterns. In land cover studies, the random forest classifier has been found to be stable, requiring very few user-defined parameters and to yield high overall accuracies (Senf et al. 2012). An overview of the methods is shown in Figure 1.

The random forest classifier was used to produce 13 annual land cover classification maps from 2001 to 2013. After implementing, the trajectory of some pixels of results may indicate unnatural behavior due to the error, for example, from agriculture to urban, then back to agriculture. To amend this, we adapted a temporal moving filter according to Clark et al. (2010). This filter corrects the land cover change maps that if the land cover is same at year $i-1$ and $i+1$, then the cover at year $i$ should be same. Finally, Accuracy assessment is implemented by 50% of reference data, which are totally 1,024 samples.

## 4. Expected results and discussion

The results of this analysis are annual land cover maps. In contrast to the MCD12Q1 product produced using a supervised classification algorithm, which is officially published as annual land cover data by MODIS Land Team, the land cover data from this analysis include measures of land cover change. By quantifying annual change in urban areas, this analysis supports a deeper understanding the urbanization process, in terms of spatial and temporal contexts. Using high frequency observations to produce land cover change information improves understanding of the temporal processes of spatial phenomena. In this case the quantification of change in urban extent is essential for researchers, urban planners, and policy makers to better understand responses in urban environments to planning and policy decisions. Critically, such detailed analyses of change also provide an evidence base to investigate non-policy related changes that may be driven by other social processes.

There are a number of items for critical discussion: i) the classification scheme, especially the number of legend categories and coarse spatial resolution, ii) the influence of climate change on changes in EVI annual data, and iii) the validation scheme for the spatio-temporal results. Due to the coarse resolution of MODIS pixel compared the high resolution satellite images such as Landsat and SPOT, the interpretation of land cover may not be straightforward. Thus, the number of classes was in this case determined through trial and error. Although this study set 3 categories of land cover which dominates more than 70% in pixel, further analysis will include fuzzy classification to allow for sub-pixel mixtures of legend categories. Additionally it is recognized that the climatic impact of temperature and precipitation on EVI time series data is inevitable even on impervious surface, because pixels usually include sub-pixel mixtures of vegetated land in the study area. Thus, some of the land cover changes identified may be as a result of changes in climate and associated responses in the EVI data. Future work will examine the influence of such environmental factors on EVI patterns particularly within periods of high temporal variability such as unusual and/or unseasonal drought, typhoons and floods. This will allow the land cover change results to be parameterized in order to accommodate any impacts of change in climate and to better

compare time-series data on change spatially and inter-annually. The validation of spatio-temporal modeling needs to be considered. Although Google Earth is one of the most useful databases which stores time-series VHR imagery to identify the actual land cover at several time steps, there is limited historical image data availability (early 2000s). Thus, the temporally bias in the availability of reference data might affect the results. Finally, a deeper understanding of the patterns of urban expansion and any errors will be developed through spatial analyses of change using spatially explicit models such as those proposed by Comber (2013).

## Acknowledgements

This research is funded by "International Network hub for Future Earth: Research for Global Sustainability" of "The Strategic Young Researcher Overseas Visit Program for Enhancing Academic Collaboration, JSPS, Japan".

## References

Clark, M. L., Aide, T. M., Grau, H. R., & Riner, G. (2010). A scalable approach to mapping annual land cover at 250 m using MODIS time series data: A case study in the Dry Chaco ecoregion of South America. *Remote Sensing of Environment, 114(11)*: 2816–2832.

Comber A.J., (2013). Geographically weighted methods for estimating local surfaces of overall, user and producer accuracies. *Remote Sensing Letters*, 4(4): 373–380.

Jönsson, P., & Eklundh, L. (2004). TIMESAT—a program for analyzing time-series of satellite sensor data. *Computers & Geosciences, 30(8)*: 833–845.

Huete, A., Didan, K., Miura, T., Rodriguez, E. ., Gao, X., & Ferreira, L. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1-2): 195–213.

Senf, C., Hostert, P., & Linden, S. Van Der. (2012). Using MODIS time series and random forests classification for mapping land use in South-East Asia. In *Geoscience and Remote Sensing Symposium (IGARSS),* IEEE International, 6733–6736.

Sexton, J. O., Song, X.-P., Huang, C., Channan, S., Baker, M. E., & Townshend, J. R. (2013). Urban growth of the Washington, D.C.–Baltimore, MD metropolitan region from 1984 to 2010 by annual, Landsat-based estimates of impervious cover. *Remote Sensing of Environment*, *129*: 42–53.

Yao, X. (2003). Research Issues in Spatio-temporal Data Mining. In *the University Consortium for Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery, Nov. 18-20,* Lansdowne, Virginia, 1–6.

# A New Approach To Cluster Validation in Experimental Investigations of (Geo)Spatial Concepts

J. O. Wallgrün[1], A. Klippel[1], D. Mark[2]

[1]The Pennsylvania State University, University Park, PA 16802
Email: {wallgrun; klippel}@psu.edu

[2]NCGIA & Department of Geography, University at Buffalo, Buffalo, NY 14228
Email: dmark@buffalo.edu

## 1. Introduction

Cluster analysis is a popular method across many disciplines and frequently employed in the spatial sciences to analyze observed or experimentally collected data. Cluster analysis either operates on entities in an *m*-dimensional feature space based on some distance measure (e.g., Euclidean distance) representing the dissimilarity of features, or, alternatively, directly on a given proximity matrix representing the similarity / dissimilarity between pairs of entities. Cluster analysis approaches can be distinguished into partitioning and hierarchical methods. While partitioning cluster methods identify cluster membership at a single level, e.g., on the basis of a predefined number of clusters, hierarchical methods iteratively join entities and clusters based on different algorithms resulting in a tree structure (dendrogram).

One of the greatest challenges of hierarchical cluster analysis is to decide how to interpret the clustering process and dendrogram, i.e., deciding on the right/best number of clusters (cluster validation). This problem is aggravated by different clustering methods being available which potentially offer different interpretations on both the number of clusters and the cluster-membership of entities.

This paper advances cluster validation for grouping experiments (category construction / free classification) in which participants organize stimuli on the basis of perceived similarity. Such experiments typically aim at shedding light on cognitive concepts and are applied in many areas, from human-computer-interaction design (e.g., Roth et al. 2011) to the assessment of qualitative spatial calculi (Mark and Egenhofer 1994, Knauff et al. 1997, Klippel et al. 2013). We propose a novel cluster validation method that determines a) the best cluster solution; b) the required number of participants, and c) enables (meta) comparisons across different experiments. We discuss results of applying this method to data collected in several geospatial behavioral studies.

## 2. Cluster Validation

Three cluster validation approaches are prominent in the literature: First, comparing the results of different clustering methods. This approach is sometimes referred to as confirmatory cluster analysis (Fisher and Ransom 1995). Reanalyzing the data using different methods can determine the extent to which solutions converge.

Second, indices (e.g., Rand Statistics, Jaccard Coefficient; see Halkidi et al. 2002) are used to assess to which degree two partitions (i.e., sets of clusters) *G* and *H*, match. These indices allow for comparing the clustering of stimuli resulting from a grouping experiment with an assumed theoretical partition or to assess how similar the results of different clustering methods are. They are based on the number of pairs of stimuli that belong to the same group in both *G* and *H* (*SS*), the number of pairs that belong to the same group in *G* but different groups in *H* (*SD*), the number of pairs that belong to the same group in *H* but

different groups in *G* (*DS*), and the number of pairs that belong to different groups in both *G* and *H* (*DD*). The Jaccard Coefficient, for instance, is computed as

$$J(G,H) = \frac{SS}{SS+SD+DS}$$

Third, if the sample is large enough, it can be randomly split and each half can be analyzed separately and solutions can be compared (Mandara 2003).

## 3. Sampling-based Cross-Method Validation Approach

Combining and extending the above mentioned validation approaches, our cluster validation method described in the following is based on the general idea of sampling from the pool of participants and computing a similarity index called *cross-method similarity index* (CMSI) for the groups obtained for a given number of clusters by applying three different clustering methods: Ward's method, average linkage, and complete linkage. The CMSI value is computed for sample sizes up to the number of available participants and different numbers of clusters *c*. Results are plotted (see Figure 4) which allows for addressing the issues described above.

The CMSI is computed as follows: Participant pool *P* consists of *m* participants $p_1, \ldots p_m$ and the grouping created by participant $p_i$ is given by an individual similarity matrix (ISM) $ISM_i$ which contains a '1' for each pair of stimuli put into the same group by $p_i$ and a '0' for pairs put into different groups.

Given the sample size *n*, a random sample $S_n$ of *n* participants is drawn from *P*. An overall similarity matrix (OSM) is then computed by adding up the ISMs for the participants contained in the sample (Figure 1). Three cluster analyses using three different methods (see above) are performed on the OSM.

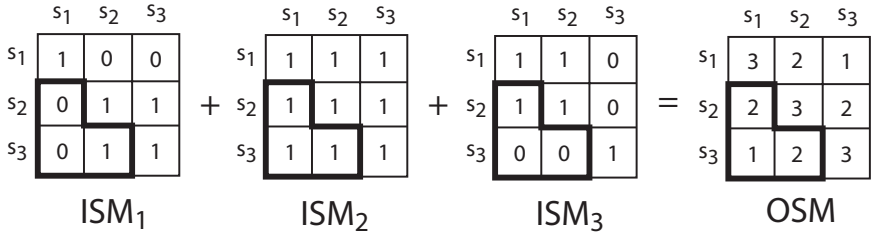



Figure 1. Computation of the OSM matrix.

The output of the clustering methods are dendrograms (see Figure 2). For a given number of clusters *c*, the respective groupings are derived which corresponds to cutting the dendrograms at a particular height (Figure 2). The resulting groupings are referred to as $G_{c,S_n}^{[ward]}$, $G_{c,S_n}^{[avg]}$, and $G_{c,S_n}^{[comp]}$. Finally, the Jaccard Coefficient is used to compute the similarity of the groupings for each pair of methods; the average is taken as the CMSI value:

$$CMSI_{c,S_n} = \frac{J(G_{c,S_n}^{[ward]},G_{c,S_n}^{[avg]}) + J(G_{c,S_n}^{[ward]},G_{c,S_n}^{[comp]}) + J(G_{c,S_n}^{[avg]},G_{c,S_n}^{[comp]})}{3}$$

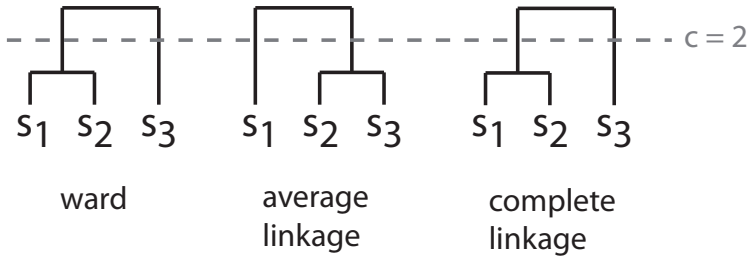



Figure 2. Dendrograms resulting from different clustering methods and cutting them to get groupings with c=2 clusters.

The CMSI value provides a measure of how well the different clustering methods agree for a given sample and number of clusters. It can easily be generalized to include other clustering methods as well. Computing the average CMSI value over many samples and for different numbers *n* and *c* and plotting it as in Figure 4 allows for a statistical perspective on the grouping behavior.

## 4. Results

We used the CMSI approach to analyze and compare the results from several grouping experiments we conducted in the past on human conceptualizations of spatial relations. We here only provide a demonstration by comparing two experiments. The first used icons showing different topological relations based on Galton's overlap relations (Galton 1998); the second used icons showing two airplanes in different direction relations (see Figure 3).

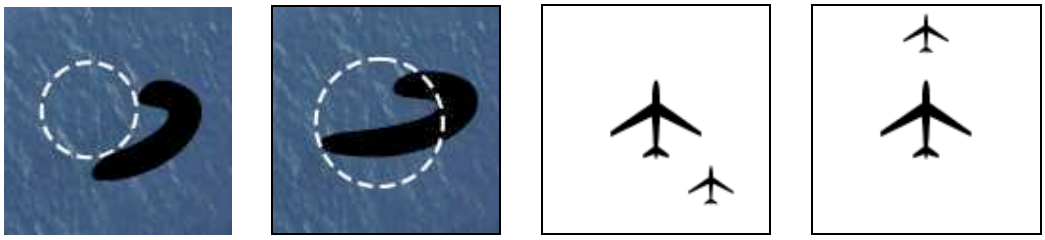


Figure 3. Icons from the two grouping experiments (left: a protected habitat in relation to an oil spill; right: two airplanes).

Figure 4 shows the resulting CMSI plots for these two experiments. In the Overlap experiment even a small numbers of participants reached the optimal CMSI score for a three-cluster solution. This corroborated findings reported in Wallgrün et al. (2013) that a three-cluster solution that separates non-overlapping, overlapping, and proper-part relations is the cognitively adequate model for modes of overlap.

In contrast, the CMSI plot for the Directions experiment does not show a perfect score at all. The conclusion is that participants adopted competing and mutually exclusive direction concepts: some used half planes and separated directions into quadrants, other used a cone shape approach. However, splitting the participants according to their strategies and applying the CMSI approach to these subgroups results in perfect scores.

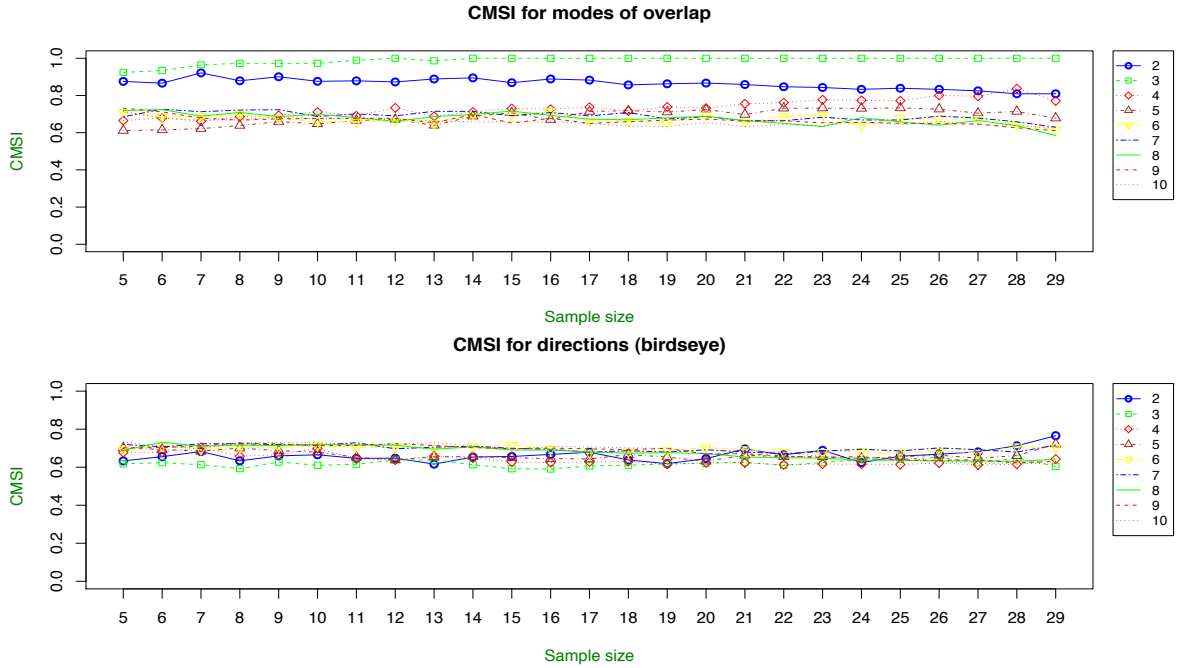



Figure 4. Plots of CMSI values for cluster numbers c=2 to 10 (see legends on the right).

## 5. Conclusions

The sampling based cross-method similarity validation method proposed in this paper is a tool to support the statistical analysis of experimental data from behavioral studies on human spatial cognition. It has the potential to reveal whether a common conceptualization of the stimuli exists or whether there is a need to acknowledge competing perspectives—crucial questions in cognition and ontology engineering. The CMSI is applicable in other domains where cluster analysis plays a role. In addition to what we could show in this abstract, we have applied it to a much larger number of experiments and to subgroups of participants resulting from a participant similarity analysis to validate clustering results.

## Acknowledgements

This research is funded by the National Science Foundation under grant #0924534.

## References

Fisher, L and Ransom, D C, 1995, An empirically derived typology of families: I. Relationships with adult health. *Family Process*, 34(2):161–182.

Galton, A, 1998, Modes of overlap. *Journal of Visual Languages and Computing, 9*:61–79

Halkidi, M, Batistakis, Y, and Vazirgiannis, M, 2002, Cluster validity methods: Part I. *ACM SIGMOD Record, 31*(2):40–45.

Klippel, A, Li, R, Yang, J, Hardisty, F, and Xu, S, 2013, The Egenhofer-Cohn hypothesis or, topological relativity? In Raubal, M, Frank, A, and Mark, D (eds.), *Cognitive and Linguistic Aspects of Geographic Space - New Perspectives on Geographic Information Research*, pp. 195–215.

Knauff, M, Rauh, R, Renz, J, 1997, A cognitive assessment of topological spatial relations: Results from an empirical investigation. In: Hirtle, Frank (eds.) *Spatial Information Theory*, pp. 193-206.

Mandara, J, 2003, The typological approach in child and family psychology: A review of theory, methods, and research. *Clinical Child and Family Psychology Review*, 6(2):129–146.

Mark, D M and Egenhofer, M J, 1994, Calibrating the meanings of spatial predicates from natural language: Line-region relations. In: Waugh, Healey, (eds.) *Advances in GIS Research*, pp. 538-553.

Roth, R E, Finch, B G, Blanford, J I, Klippel, A, Robinson, A C, and MacEachren, A M, 2011, The card sorting method for map symbol design. *Cartography and Geographic Information Science, 38*(2):89–99.

Wallgrün, J O, Yang, J, Klippel, A, 2013, Investigating intuitive granularities of overlap relations. In *12th IEEE International Conference on Cognitive Informatics and Cognitive Computing*, 2013.

# Evaluating minerals deposits prospectivity using multisource data based on fuzzy decision method in GIS

Ping Wang[1], Xiangnan Liu[2], Meiling Liu[2], Qin Yang[2]

[1]School of geographical science, Northeast Normal University,Changchun, 130024, China
Email:wangp666@nenu.edu.cn

[2]School of Information Engineering, China University of Geosciences, Beijing 100083, China
Email: liuxncugb@163.com

## 1. Introduction

A major challenge for mineral exploration geologists is the development of a transparent and reproducible approach to targeting exploration efforts.

The objective of this research is to identify minerals deposits prospectivity by combining GIS with fuzzy decision method based on multi-source (i.e., geological, geochemical and remote sensing data).

The assessment of Cu-Zn deposits prospectivity located in Yushigou, Qinghai, China was used as a case study. The study area (38°35′N to 38°40′N, 98°30′E to 99°00′E) is shown in Figure1. It is within plateau-climate, a high attitude and poor vegetation cover (Zhao and Chen, 1999). Most of the geological units in the study area have been altered hydrothermally and exposed with abundant faulting tectonics. It locates at northeastern margin of the Qinghai-Tibet plateau and is bounded to the south by the Qaidam basin and to the north by the Tarim basin and Alxa block. Rock masses are mainly composed of the ophiolite suite, the cumulate dunite, gabbro, troctolite, diabase and basalt rock with well differentiation, obvious facies, and strong tectonic destruction, where is generally broken rock, widely variable low-temperature hydrothermal alteration in the late stage.

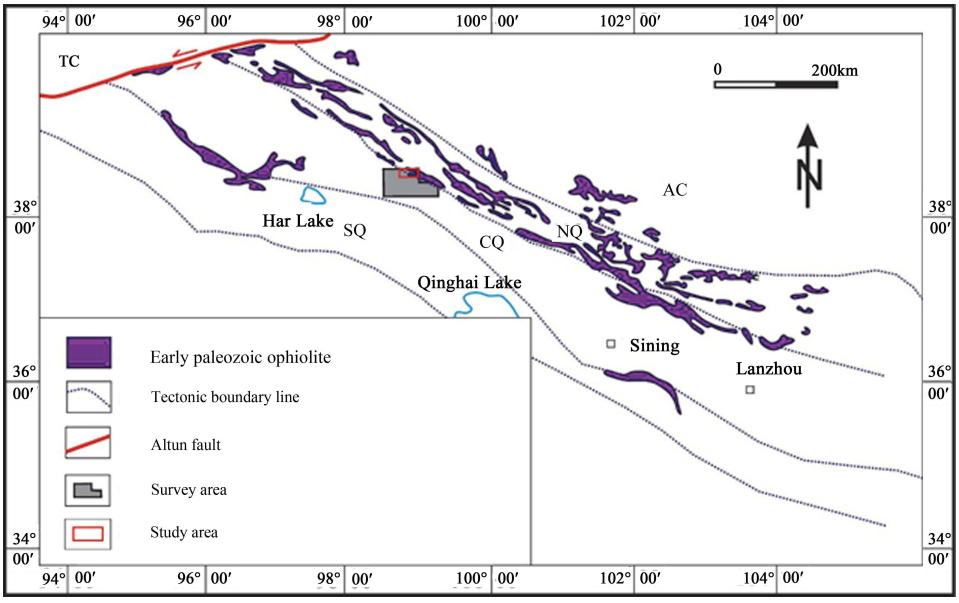



Figure 1. Tectonic of Qilian Mountains

AC- Alxa continental block; NQ- North Qilian suture zone； CQ- Centre of Qilian continental block；SQ- South Qilian continental block; TC- Tarim continental block

# 2. Data and method

## 2.1 Data

Here, multi-source data sets involving geological, geochemical and remote sensing data were collected in this study. Namely, foundation geological map (1:500, 00 scale), geochemical data in 1980s (1:200,000 scale) and ASTER data on August 2004 were obtained.

## 2.2 Method

GIS in combination with fuzzy decision method is adopted to map mineral deposits prospectivity. The spatial analysis of GIS technology, such as overlay, buffer, correlation, interpolate and logical method, play an important role in extracting the factors from the multi-information of geology, geochemistry, and remote sensing in mineral prediction and assessment (Zhou et al., 2007). The compromise fuzzy multiple attribute decision making (CFMADM) as one of fuzzy decision methods, which develops from classical multi-criteria fuzzy multiple attribute decision making, was selected (Zuo et al.,2009; Adiat et al.,2012; Abedi, 2012). The procedure for the assessment of minerals deposits prospectivity using CFMADM can be summarized as follows. (1) Fuzzy positive-ideal index and fuzzy negative-ideal index was displayed from original sample data. They are composed of maximum fuzzy index and minimum of fuzzy index, respectively. (2)The distance between each object and the fuzzy positive-ideal index, between each object and fuzzy negative-ideal index are calculated by weighted euclidean distance. (3) Relative-membership grade of each object which belongs to the fuzzy positive-ideal is calculated based on interpretation of mineralization possibility in each area. The higher the membership degree is, the greater the indication of an existing possibility mineralization.

Various evidential layers from different multi-source data sets are utilized to assess minerals deposits prospectivity. Hydrothermal alteration zones were extracted using Principal Component Analysis (PCA) technology from ASTER data (Gillespie et al., 1986; Shi et al., 2012). Namely, ASTER bands 1, 2, 3 and 4 are applied in PCA for silicification alteration information and band 1, 3, 4 and 8 are applied in PCA for oxhydryl alteration information. Ore-controlling faults and host rock areas were extracted from the foundation geological map. Four buffer zones along the faults, namely <250 m, 250-500 m, 500-750 m, >750 m buffer were established, which represent the presence of mineralization beneath the adjacent faults. Copper anomaly and zinc anomaly were interpolated using an inverse distance weighted from geochemical data.

# 3. Result

## 3.1 Evidential layers extraction

Six indies are considered as key factors to assess Cu-Zn deposits prospectivity, namely, Silicification, Hydroxyl alteration, fault classification, formation lithology classification, copper anomaly and zinc anomaly (Figure 2). Cell values of raster data associated with these criteria are extracted and stored in 6 separate columns and 3219 rows in a database (Table 1). Four category in each layer was classified, namely non-anomalous (D), possibly anomalous(C), probably anomalous (B) and anomalous (A). The normal or log-normal distribution of data such as alteration layers and chemical anomalies layers can be reclassified by its standard deviation (σ) and mean ( $\tilde{x}$ ) into ( $\tilde{x}+\sigma$ ), ( $\tilde{x}+2\sigma$ ) and ( $\tilde{x}+3\sigma$ ) classes. Values less than ( $\tilde{x}+\sigma$ ) was classified as D degreed and values more than ( $\tilde{x}+3\sigma$ ) was considered as

A. The value of ( $\tilde{x}+2\sigma$ ) and ( $\tilde{x}+3\sigma$ ) are thresholds to define possibly and probably anomalous areas. In addition, fault layer is classified into four classes (A, B, C and D) associating with the buffer map, where the buffer area closest to the fault is assigned to A, the secondary ones assigned to B, and others are assigned in proper sequence. In particular, host rock layer is classified into two classes, where there is the host rock such as peridotite diabase and basalt units is considered as A, and other areas is considered as D due to its lacking in mineralization . Weights for each evidence layers were established by a Delphi method, a knowledge-driven method, and asking questions by a group of experts in this field.

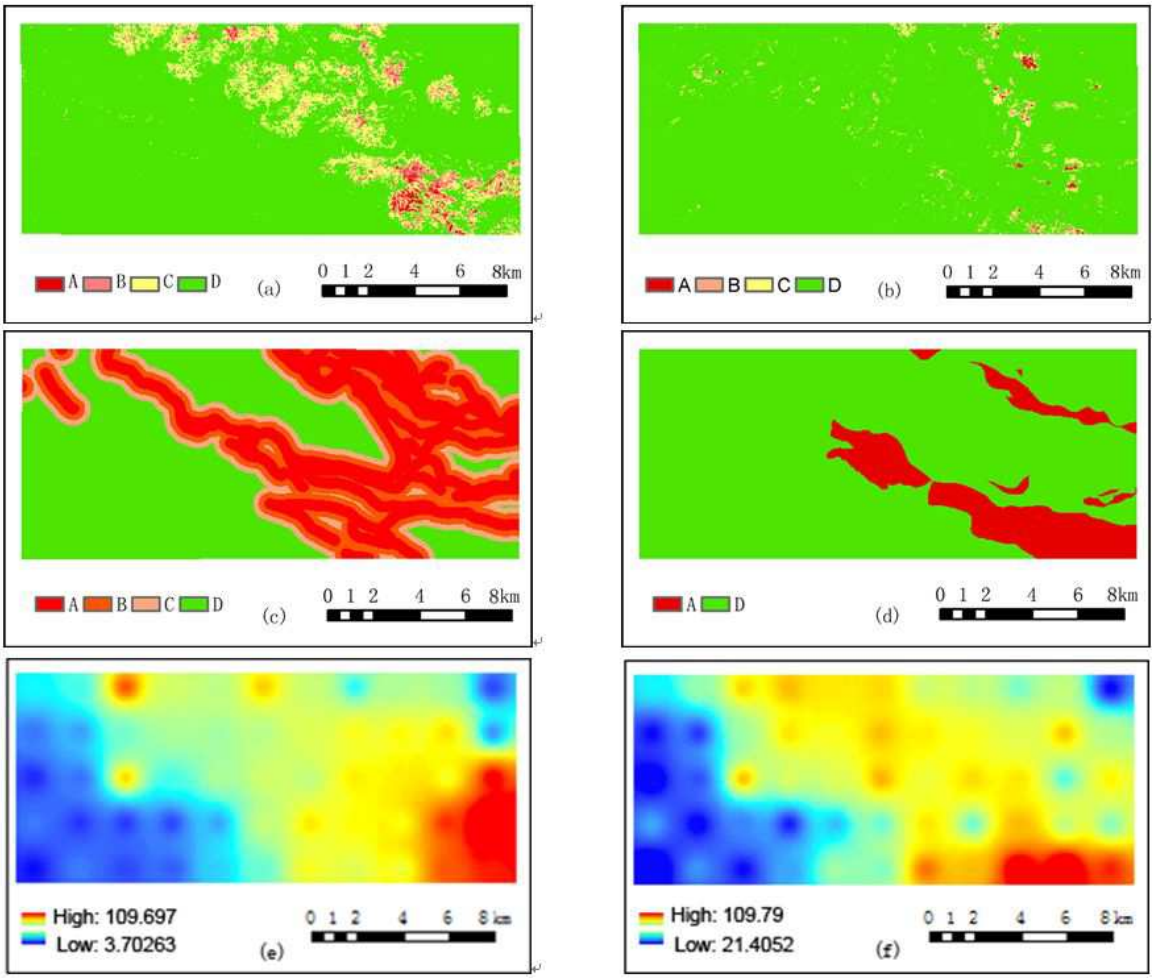



Figure 2 Six indies for Ore control factors the Yushigou area,,Qilian Mountains, Qinghai: (a) Silicification classification map information; (b) Hydroxyl alteration classification; (c) .Fault classfication; (d) Formation lithology classfication;(e)Copper anomaly distribution (f) Zinc anomaly distribution

Table 1 Six indies for Ore control factors by extracting of one evaluation cell in the Yushigou area,Qilian Mountains.

| Id | Silicification | Hydroxyl alteration | Fault | Formation lithology | Copper | Zinc |
|---|---|---|---|---|---|---|
| 1 | A | A | D | D | 6.60474 | 27.0473 |
| 2 | C | D | C | B | 5.93221 | 25.8726 |
| 3 | A | A | D | D | 5.54322 | 25.1072 |
| 4 | B | C | D | C | 5.65065 | 25.3239 |
| … | … | … | … | … | … | … |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3219 | D | A | B | D | 6.39017 | 26.8113 |

### 3.2 Minerals deposits prospectivity mapping

The comprehensive membership is obtained by applying CFMADM to all evidential layers. All cells value is transformed to triangle fuzzy numbers, then, fuzzy decision matrix is built by normalization and weighted method for the fuzzy numbers. The final comprehensive membership was computed using CFMADM implemented in Matlab. The result showed that membership values are concentrated on 0.34-0.78. In addition, the best threshold is set as 0.6 through experimental test based on the distribution feature of membership and ore-controlling factors. Values high 0.7 is taken as the best probability mineralization, values between 0.6 and 0.7 is viewed as the possibly mineralization, while values lower 0.6 is the candidate possibly mineralization area. As a result, 8 mine zones for copper and zinc has been assessed (Figure 3).

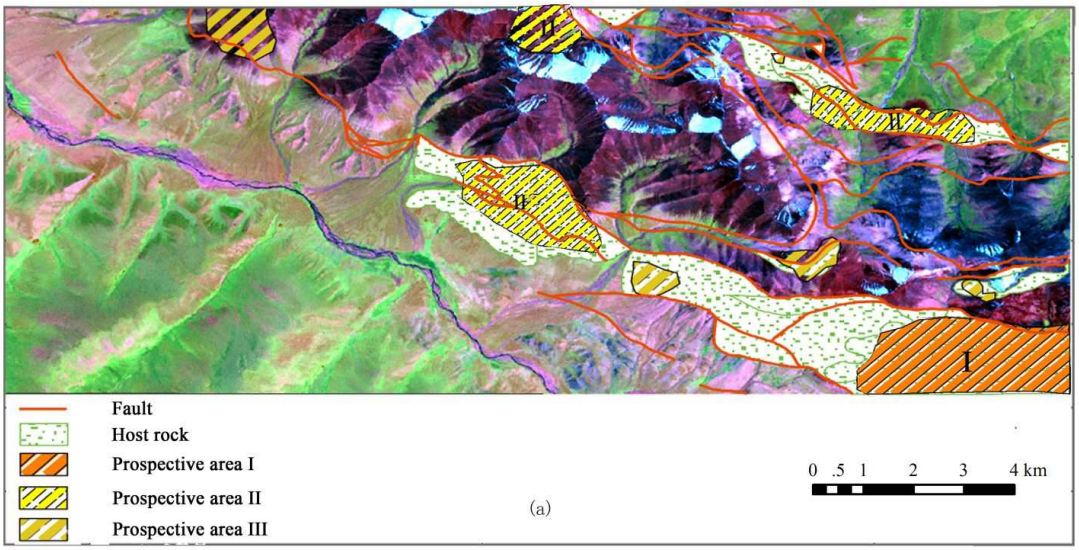



Figure 3 Mapping of Cu-Zn deposits prospectivity in the Yushigou area, Qilian Mountains by using GIS and fuzzy decision method

## 4. Conclusion

Distribution of prospective area basically reflects the main metallogenic laws of Cu-Zn ore. The main mineralization is located in magmatite intrusion stratum contacts and among the main fault zone. The most prospecting areas is located in the junction of Tuolaishan deep fault, where there is exposed with massive basalt tuff, and a favorable area for prospecting. The three prospecting areas for secondary probably is distributed around the most probability prospecting areas and closed to the Tuolaishan deep fault, where there are ultrabasic bodies or basic bodies contact zone. It indicates that good ore prospects are close related to the strong development of fracture convergence zone.It concluded that CFMADM and GIS technology are useful for prospecting minerals deposits and for identifying new exploration targets.

## Acknowledgements

This research was supported by the project “Survey and Evaluation of Geology and Mineral Resources of Metallogenic belt at the peripheral and adjacent area of Qaidam Basin, Qinghai province”of China Geological Survey (project ID: 1212011121188).

## References

Abedi M, 2012. Fuzzy outranking approach: A knowledge-driven method for mineralprospectivity mapping. *International Journal of Applied Earth Observation and Geoinformation*, 23,1-12.

Adiat KAN, Nawawi MNM and Abdullah K, 2012. Assessing the accuracy of GIS-based elementary multi criteria decision analysis as a spatial prediction tool – A case of predicting potential zones of sustainable groundwater resources. *Journal of Hydrology*, 440-441,75-89.

Gillespie A, Kahle AB, Walker RE, 1986. Color enhancement of highly correlated images. I. Decorrelation and HSI contrast stretches. *Remote Sensing of Environment,* 20, 209–235.

Shi PL, Fu BH, Ninomiya Y, Sun JM and Li Y, 2012, Multispectral remote sensing mapping for hydrocarbon seepage-induced lithologic anomalies in the Kuqa foreland basin, south Tian Shan. *Journal of Asian Earth Sciences*, 46,70-77.

Zhao J and Chen CK, 1999. Geography of China. Higher Education Press, Beijing 606–615.

Zhou_W, Chen G, Li H, Luo HY and Huang SL,2007,GIS application in mineral resource analysis—A case study of offshore marine placer gold at Nome, Alaska. *Computers & Geosciences*, 33,773–788.

Zuo RG, Cheng QM and Frederik P, 2009, Application of a hydrid method combining multilevel fuzzy comprehensive evaluation with asymmetric fuzzy relation analysis to mapping prospectivity. *Ore Geology Reviews,*35,101-108.

# Development of Social Media GIS in Order to Accumulate, Share and Exchange Regional Information

Kayoko Yamamoto[1]

[1]University of Electro-Communications, 1-5-1 Chofugaoka Chofu-shi Tokyo 182-8585 Japan
Email: k-yamamoto@is.uec.ac.jp

## 1. Introduction

In recent years in Japan, where the formation of a highly information-oriented society is being achieved, the amount of information about urban areas is on the increase, and a diverse range of information can easily be obtained by various means anywhere, anytime. However, in regions outside urban areas, although the amount of information is increasing, compared with urban areas, it can by no means be termed sufficient. Further, it is difficult for people other than those who reside, commute to work, or attend schools in the regions outside urban areas to obtain and utilize detailed regional information. Taking these background factors into account, this study aims to develop a social media GIS which enables accumulation and sharing of regional information and exchange of information between regions, in order to supplement the scarcity of information in regions outside urban areas.

## 2. System Design

In this system, as shown in Figure 1, three web applications, that is, a Web-GIS, a SNS and Twitter, were integrated to develop a social media GIS that is effective for information exchange between regions which is based on the accumulation and sharing of regional information. The method for integrating these three web applications was to include the Web-GIS in the SNS, and conduct a mashup using the SNS and Twitter. The system enables geographical understanding of location information relating to information contributed, via the Web-GIS; management and visualization of information contributed on the digital map which includes environment variables; accumulation and sharing of regional information of users and exchange of information between regions using the self-developed SNS; and classification of the importance of contributed information. Further, by enabling the contribution of information from Twitter as well, user stress is relieved and long-term operation is realized; further, users inside and outside the region of operation can use Twitter to easily contribute information from a portable information terminal anytime, regardless of whether they are indoors or outdoors.

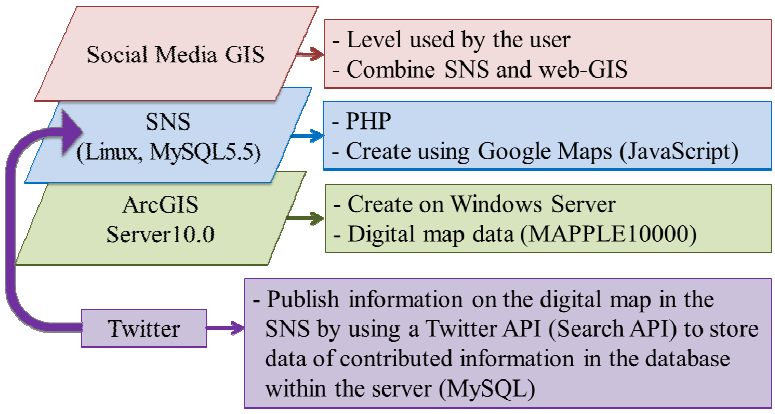



Figure 1. System design

# 3. System Development

## 3.1 System Front End

**(1)Personal data registration/profile publication functions**

The first time a user makes access to the system, they use the initial registration screen to register personal data such as their “User ID”, “Password”, “Age group”, “Sex”, “Region” and “Greeting”. This is because it is desirable that the system should be designed such that when users conduct interactive communication with each other, they can be identified to a certain extent. Further, because users who do not wish to make their personal data public have been taken into consideration, the “User ID” is designed such that the user’s real name and account name are not specified, and the user can freely select and enter a user ID.

**(2)Information contribution/browsing functions**

Two types of methods were provided for when a user contributes regional information – the method of contributing information from a computer, and the method of contributing information from a portable information terminal using Twitter. In the former method, first, the user clicks “Post” on the home page of the website on their computer screen, to go to the posting page. On the posting page there is a form in which the user can enter the “Title” and “Main text”. After the user has entered the content into the form, location information relating to the posted content can be added simply by clicking the posting location on the digital map. The “location information” is entered into MySQL, and when transmission is performed, posting is complete. In the latter method, data of information posted using Twitter from a portable information terminal is acquired, and displayed on the digital map on the posting page of the user which is set up in the system. In both methods, at the time of posting, if necessary, a photo image file can also be attached. All contributed information is shared within the SNS. From the home page, the following three sections can be browsed: The ten most recently contributed items of information, a list of contributed information, and the points ranking of contributed information.

**(3)Button functions/ranking function**

Button functions are used for classifying the importance of contributed information. Two types of buttons were provided - “I didn’t know” for users within the region of operation, and “I want to go there” for users outside the region of operation. Thus, the provision of button functions in this system enables users to easily express their intention in regard to information they have viewed. In this study, when a user outside the region of operation uses the above-mentioned comment function and the “I want to go there” button function in response to information posted by a user in the region of operation, it is defined as an exchange of information between regions. Further, for each item of contributed information, one point is added each time users inside and outside the region of operation click either of the two above-mentioned buttons, and each piece of contributed information is evaluated. Moreover, by including a ranking function which displays contributed information in descending order of total points gained, the system avoids losing contributed information which users are strongly interested in amongst other information.

## 3.2 System Back End

**(1)Management system for contributed information that is run by administrators**

Administrators log in from a login page exclusively for administrators, and a screen exclusively for administrators is provided. Using the administrator screen, administrators manage users, and in cases where there has been an inappropriate statement or inappropriate behavior, they manage the matter by taking action, such as closing user accounts. Further, administrators can view all contributed information, contributor names, and dates and times of contributions on a screen which lists them; therefore, if by any chance an appropriate

posting is made, they can delete it with just one click. Thanks to these aspects of the system, the burden of administrators can be reduced, because there is no need for them to search to check whether there are inappropriate items of contributed information in the system. Further, the case where local residents actually perform the role of administrators in the regional community is anticipated. The system is designed such that MySQL is managed using graphical user interface (GUI) and administrators who do not have a very high level of IT literacy can also manage and administer; therefore, the burden on administrators can be reduced as much as possible.

**(2)Mashup system with Twitter**

In this study, when a mashup is performed with Twitter, the Search API with Basic authentication protocol is used, and thereby the effort involved in information contribution is minimized and user stress is reduced. Conventionally, when a Twitter mashup system is developed, the OAuth authentication protocol is often used. However, in this study, upload from the main part of the system to Twitter of information for contribution and so forth is not conducted; therefore, the Search API with Basic authentication protocol, which allows acquisition by searching Twitter data, was employed. The data for reflection in the system (main text, location information, account names, dates and times) is obtained by making a query specification. In the process for acquiring data of information contributed using Twitter from a portable information terminal, users simply register a Twitter account name in the blank at the time of initial registration. The rest is performed by the back end. Data of information contributed using Twitter of registered users is obtained by applying all account names saved in the database to the "user" portion of "from:<user>" of tags.

## 3.3 System interface

The system has three types of interface – the computer screen of the user (Figure 2), the portable information terminal screen of user, and the computer screen of the administrator. Using the administrator screen, inappropriate contributed information can be promptly deleted, and any user can also make an amendment using either of the two types of user screen, by making a comment in response to erroneous contributed information. Thus, this system has been developed keeping in mind the goals of reducing administrator burden as much as possible and developing a system which regular local residents in regional communities can also operate and manage.


| No. | Description |
|---|---|
| 1 | User greeting |
| 2 | User profile publication |
| 3 | The ten most recent items of information contributed from portable information terminals using Twitter |
| 4 | Go to a list of information contributed from portable information terminals using Twitter and the ranking page |
| 5 | The ten most recent items of information contributed from computers |
| 6 | Go to the home page of the user (sample information is displayed on a digital map) |
| 7 | Go to the page which contains messages from administrators |
| 8 | Go to the page where change and registration of personal data can be made |
| 9 | Go to the page where information can be contributed from a computer |
| 10 | Go to the page where information contributed from computers can be viewed |
| 11 | Go to the page where information contributed from portable information terminals using Twitter can be viewed |

Figure 2. Illustration of User Computer Screen and Functions

## Acknowledgment

In the operation of the social media GIS, enormous cooperation was received from those in the eastern region of Yamanashi Prefecture and the Tama region of Tokyo in Japan. I would like to take this opportunity to gratefully acknowledge them.

## References

Yanagisawa T and Yamamoto K, 2012, A Study on Information Sharing GIS to Accumulate Local Knowledge in Local Communities. *Theory and Applications of GIS*, 20(1): 61-70.

Nakahara H, Yanagisawa T and Yamamoto K, 2012, A Study on a Web-GIS to Support the Communication of Regional Knowledge in Regional Communities: Focusing on Regional Residents' Experiential Knowledge, *Socio-Informatics*, 1(2):77-92.

Yamada S and Yamamoto K, 2013, Development of social media GIS for information exchange between regions. *International Journal of Advanced Computer Science and Applications*, 4(8):62-73.

# PyGWA: A Python Library for Geographically Weighted Analysis

J. Yao[1], A.S. Fotheringham[2]

[1]School of Geography & Geosciences, University of St Andrews, St Andrews, Fife KY16 9AL, Scotland, UK
Email: jing.yao@st-andrews.ac.uk

[2]School of Geographical Sciences & Urban Planning, Arizona State University, P.O. Box 875302, Tempe AZ 85287-5302, USA
Email: Stewart.Fotheringham@asu.edu

## 1. Introduction

Spatial data analysis and spatial statistical analysis in particular is a fundamental subfield of GIScience. It contains a variety of techniques using topological or geographic properties to solve spatially explicit problems. Recent years have seen a dramatic growth in the theories, methodologies as well as software implementations in spatial statistics (Anselin 2010, Goodchild 2010, Anselin and Rey 2012), which have been extensively applied in public health, housing market research, demography, ecology and criminology, among many others.

Of particular interest in this paper is the software to implement spatial statistical techniques. So far, a number of software tools have been developed and they have greatly facilitated and encouraged the adoption and use of spatial statistics methods. Some prominent examples include SpaceStat (Anselin 1991), the spatial econometrics toolbox for MATLAB (LeSage 1999), GeoDa (Anselin et al. 2006), PySAL (Rey and Anselin 2010), GWR (Fotheringham et al. 2002) and spatial analysis libraries (e.g. "spdep" and "spgwr") in R open source statistical programming framework.

Most of the above packages are concerned with global spatial models and current local spatial modeling tools are largely limited to regression analysis. The aim of this paper is to introduce a Python library for geographically weighted analysis (PyGWA), an implementation of local spatial analysis techniques which particularly accounts for spatial heterogeneity/non-stationary. In addition to the well-known geographically weighted regression (GWR), PyGWA includes several recently developed methods that are not available in existing software.

## 2. Framework and Components

PyGWA is a software library of computational tools for geographically weighted analysis written in the open source Python language. The modules in PyGWA are designed in a way such that they are relative independent but also can be linked together to carry out certain spatial analysis. Moreover, they can be flexibly integrated with other libraries (e.g. Tkinter or wxPython) within Python for customized applications or other external tools (e.g. ArcGIS) for mapping and GIS functionality. Most importantly, PyGWA has the advantage of portability across multi-platforms which is an inherent characteristic of Python language.

The classification of different components in PyGWA largely follows the steps in a typical spatial data analysis process. Table 1 illustrates the key components of PyGWA. It distinguishes between basic data analysis and local spatial modeling. The former focuses on data processing and description, and the latter stresses geographically weighted (GW) modeling.

The most relevant parts of basic data analysis to local spatial modeling are spatial weights and visualization. The spatial weights module supports spatial kernel function definition and

optimal bandwidth selection. The visualization module contains tools for both basic exploratory data analysis (e.g. scatter plot and parallel coordinate plots) and cartographical mapping, which are important ways to help formulate hypotheses and examine the spatial analysis results.

The local spatial modeling part covers three broad categories of geographically weighted models: regression, interaction and principal component analysis (PCA). The general functionality contains model estimation, diagnostics and significance tests. Besides, the robust versions of those models are implemented as well.

Table 1. Main components and functionality in PyGWA.

| Component | Functions |
|---|---|
| ***Basic Data Analysis*** | |
| File input/output | Read and write data to spatial/non-spatial data files |
| GW summary statistics | Calculation of local statistics |
| Spatial weights | Construction of spatial weights |
| Visualization | Graphic display of spatial/non-spatial data |
| | |
| ***Local Spatial Modeling*** | |
| GW Regression | Classic GWR and its variations |
| GW interaction models | GW constrained spatial interaction models |
| GW PCA | GWR constrained PCA |

As discussed above, PyGWA is primarily intended to offer computational tools for general geographically weighted modeling and analysis. Meanwhile common routines are also provided to ease the data input, manipulation and output. All the components in Table 1 thus constitute the framework of PyGWA.

## 3. Empirical illustrations

In this section, two empirical examples are presented to illustrate how the core modules in PyGWA can be utilized to carry out local spatial data analysis and furthermore how the core functions can be accessed by customized graphical user interface (GUI) for specialized analysis.

### 3.1 Spatial Weights

Spatial weights are essential in local spatial analysis as they quantify the neighborhood structure for spatial entities. Usually, spatial weights are defined by spatial kernel functions. In PyGWA, currently two types of kernel functions are supported: Gaussian and bi-square. The bandwidth as a key element of kernel functions can be either pre-specified or determined by automatic search routines. Figure 1 shows an example demonstrating how to construct spatial weights using a bi-square kernel function with a pre-determined bandwidth.

The implementation is carried out using the command line. The spatial weights are defined by the GWR_W function within the Kernel class (a concept in the object-oriented design) which contains all the necessary functions for spatial weight calculation. It also should be noted that another relevant class FileIO is called here to obtain the geographical coordinates required for spatial weight construction.

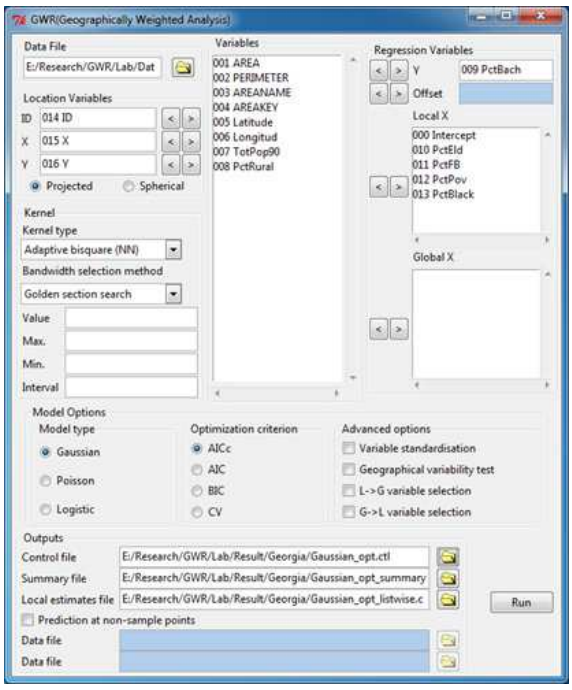
```
import FileIO

# read data
flePath = "E:/PyGWA/Sample data/Georgia/GeorgiaEduc.dbf"
allData = FileIO.read_FILE[1](flePath)

# get coordinates
flds = ['X', 'Y']  #  variables containing geographical coordinates
coords = FileIO.get_subset(allData[0], allData[1], flds)

# spatial weights definition
bdwidth =  121  # number of nearest neighbors
wType = 3  # type of kernel function: bi-square
distType = 0 # type of distance: Euclidean
weit = Kernel.GWR_W(coords, bdwidth, wType, None, distType)
```

Figure 1. Construction of spatial weights.

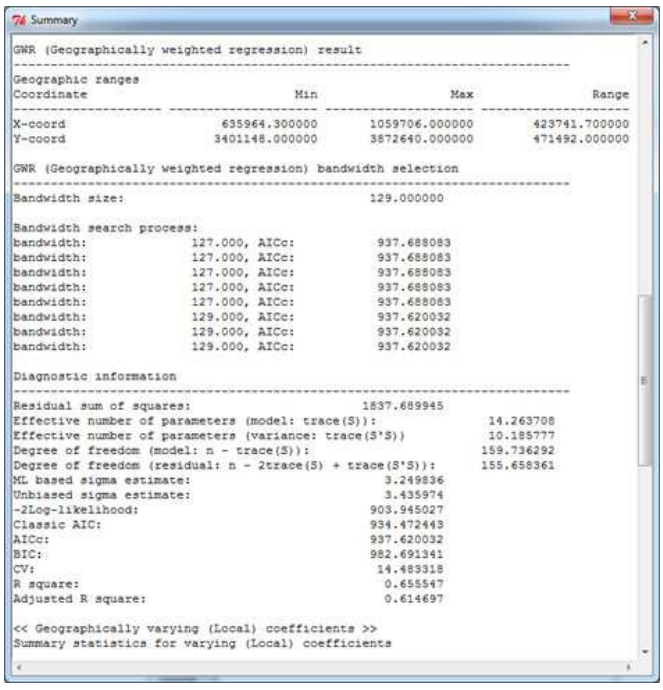
### 3.2 A Software Prototype for GWR Analysis

This section describes a software prototype customized for GWR analysis by integrating the function of GWR within PyGWA with Tkinter, a package within Python for standard GUI. Figure 2 illustrates the GUI, outputs for summary statistics and local parameter estimates, respectively.

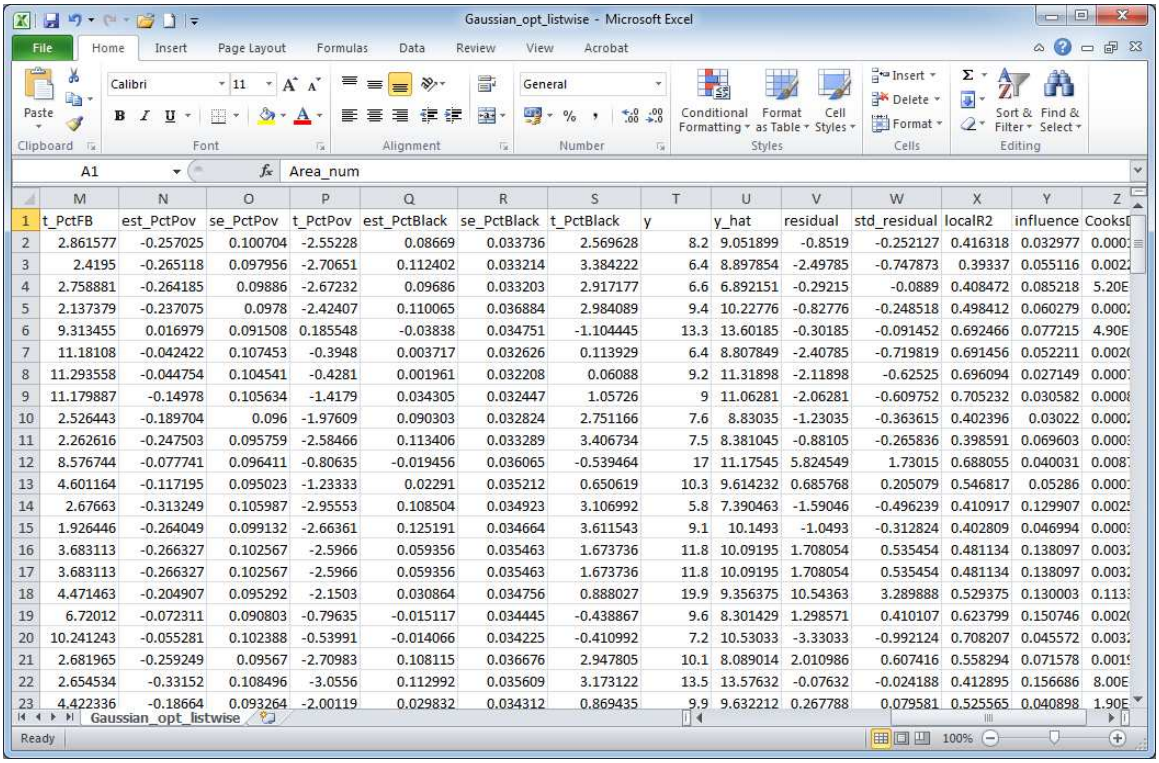
As can be seen in Figure 2a, users can build a GWR model by means of providing all the necessary parameters through the GUI, which generally includes the input file, spatial kernel function definition, model type, specification of dependent/independent variables, and model optimization criterion, etc. For example, in this case a Gaussian model is to be fitted, and the adaptive bi-square kernel function is adopted with the optimal bandwidth determined by the golden section search routine. The summary statistics of model calibration are given by Figure 2b, providing the information about model settings, bandwidth search process, diagnostics and so on. Figure 2c shows an example of the output file (.csv) which contains the parameter estimates (values, standard errors and t statistics) of all the local regression coefficients, residuals, local R square and other statistics for each observation.




(a) GUI




(b) Output of summary statistics




| | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | t_PctFB | est_PctPov | se_PctPov | t_PctPov | est_PctBlack | se_PctBlack | t_PctBlack | y | y_hat | residual | std_residual | localR2 | influence | CooksD |
| 2 | 2.861577 | -0.257025 | 0.100704 | -2.55228 | 0.08669 | 0.033736 | 2.569628 | 8.2 | 9.051899 | -0.8519 | -0.252127 | 0.416318 | 0.032977 | 0.000 |
| 3 | 2.4195 | -0.265118 | 0.097956 | -2.70651 | 0.112402 | 0.033214 | 3.384222 | 6.4 | 8.897854 | -2.49785 | -0.747873 | 0.39337 | 0.055116 | 0.002 |
| 4 | 2.758881 | -0.264185 | 0.09886 | -2.67232 | 0.09686 | 0.033203 | 2.917177 | 6.6 | 6.892151 | -0.29215 | -0.0889 | 0.408472 | 0.085218 | 5.20E |
| 5 | 2.137379 | -0.237075 | 0.0978 | -2.42407 | 0.110065 | 0.036884 | 2.984089 | 9.4 | 10.22776 | -0.82776 | -0.248518 | 0.498412 | 0.060279 | 0.000 |
| 6 | 9.313455 | 0.016979 | 0.091508 | 0.185548 | -0.03838 | 0.034751 | -1.104445 | 13.3 | 13.60185 | -0.30185 | -0.091452 | 0.692466 | 0.077215 | 4.90E |
| 7 | 11.18108 | -0.042422 | 0.107453 | -0.3948 | 0.003717 | 0.032626 | 0.113929 | 6.4 | 8.807849 | -2.40785 | -0.719819 | 0.691456 | 0.052211 | 0.002 |
| 8 | 11.293558 | -0.044754 | 0.104541 | -0.4281 | 0.001961 | 0.032208 | 0.06088 | 9.2 | 11.31898 | -2.11898 | -0.62525 | 0.696094 | 0.027149 | 0.000 |
| 9 | 11.179887 | -0.14978 | 0.105634 | -1.4179 | 0.034305 | 0.032447 | 1.05726 | 9 | 11.06281 | -2.06281 | -0.609752 | 0.705232 | 0.030582 | 0.000 |
| 10 | 2.526443 | -0.189704 | 0.096 | -1.97609 | 0.090303 | 0.032824 | 2.751166 | 7.6 | 8.83035 | -1.23035 | -0.363615 | 0.402396 | 0.03022 | 0.000 |
| 11 | 2.262616 | -0.247503 | 0.095759 | -2.58466 | 0.113406 | 0.033289 | 3.406734 | 7.5 | 8.381045 | -0.88105 | -0.265836 | 0.398591 | 0.069603 | 0.000 |
| 12 | 8.576744 | -0.077741 | 0.096411 | -0.80635 | -0.019456 | 0.036065 | -0.539464 | 17 | 11.17545 | 5.824549 | 1.73015 | 0.688055 | 0.040031 | 0.008 |
| 13 | 4.601164 | -0.117195 | 0.095023 | -1.23333 | 0.02291 | 0.035212 | 0.650619 | 10.3 | 9.614232 | 0.685768 | 0.205079 | 0.546817 | 0.05286 | 0.000 |
| 14 | 2.67663 | -0.313249 | 0.105987 | -2.95553 | 0.108504 | 0.034923 | 3.106992 | 5.8 | 7.390463 | -1.59046 | -0.496239 | 0.410917 | 0.129907 | 0.002 |
| 15 | 1.926446 | -0.264049 | 0.099132 | -2.66361 | 0.125191 | 0.034664 | 3.611543 | 9.1 | 10.1493 | -1.0493 | -0.312824 | 0.402809 | 0.046994 | 0.000 |
| 16 | 3.683113 | -0.266327 | 0.102567 | -2.5966 | 0.059356 | 0.035463 | 1.673736 | 11.8 | 10.09195 | 1.708054 | 0.535454 | 0.481134 | 0.138097 | 0.003 |
| 17 | 3.683113 | -0.266327 | 0.102567 | -2.5966 | 0.059356 | 0.035463 | 1.673736 | 11.8 | 10.09195 | 1.708054 | 0.535454 | 0.481134 | 0.138097 | 0.003 |
| 18 | 4.471463 | -0.204907 | 0.095292 | -2.1503 | 0.030864 | 0.034756 | 0.888027 | 19.9 | 9.356375 | 10.54363 | 3.289888 | 0.529375 | 0.130003 | 0.113 |
| 19 | 6.72012 | -0.072311 | 0.090803 | -0.79635 | -0.015117 | 0.034445 | -0.438867 | 9.6 | 8.301429 | 1.298571 | 0.410107 | 0.623799 | 0.150746 | 0.002 |
| 20 | 10.241243 | -0.055281 | 0.102388 | -0.53991 | -0.014066 | 0.034225 | -0.410992 | 7.2 | 10.53033 | -3.33033 | -0.992124 | 0.708207 | 0.045572 | 0.003 |
| 21 | 2.681965 | -0.259249 | 0.09567 | -2.70983 | 0.108115 | 0.036676 | 2.947805 | 10.1 | 8.089014 | 2.010986 | 0.607416 | 0.558294 | 0.071578 | 0.001 |
| 22 | 2.654534 | -0.33152 | 0.108496 | -3.0556 | 0.112992 | 0.035609 | 3.173122 | 13.5 | 13.57632 | -0.07632 | -0.024188 | 0.412895 | 0.156686 | 8.00E |
| 23 | 4.422336 | -0.18664 | 0.093264 | -2.00119 | 0.029832 | 0.034312 | 0.869435 | 9.9 | 9.632212 | 0.267788 | 0.079581 | 0.525565 | 0.040898 | 1.90E |

(c) Output of local estimates

Figure 2. Standalone software for GWR.

## 4. Summary

This paper presents the design, components and use of a software library for geographically weighted analysis, PyGWA. It will contribute to the growing domain of spatial statistical analysis software and benefit the research work in both GISciences and other disciplines. Compared to current local spatial analysis tools such as GWR4.0, GWR tool in ArcGIS, and R packages "spgwr", "gwrr" and "fdgwr" which solely focus on regression analysis, PyGWA implements a variety of both traditional and novel local analysis techniques and therefore supports more general geographically weighted analysis. Regarding future work, in one respect, we will continue improving current components of the library with an emphasis on the computational efficiency. In another respect, we will incorporate new developments in local spatial analysis into the functional modules within PyGWA. We expect to release PyGWA as well as the software prototypes built upon it in the near future.

## References


Anselin L, 1991, *SpaceStat, a software program for analysis of spatial data*. Santa Barbara: National Center for Geographic Information and Analysis (NCGIA), University of California, Technical report.

Anselin L, 2010, Thirty years of spatial econometrics. *Papers in Regional Science*, 89 (1):3–25.

Anselin L and Rey SJ, 2012, Spatial econometrics in an age of CyberGIScience. *International Journal of Geographical Information Science*, 26(12):2211–2226.

Anselin L, Syabri I and Kho Y, 2006, Geoda: an introduction to spatial data analysis. *Geographical Analysis*, 38(1):5–22.

Fotheringham AS, Brunsdon C and Charlton M, 2002, *Geographically Weighted Regression: the analysis of spatially varying relationships*. Wiley, Chichester.

Goodchild MF, 2010, Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science*, 1:3–20.

LeSage JP, 1999, *Applied econometrics using MATLAB*. University of Toledo, Technical report.

Rey SJ and Anselin L, 2010, PySAL: a Python library of spatial analytical methods. In: Fischer MM and Getis A (eds), *Handbook of applied spatial analysis*, Berlin: Springer, 175–193.

# The Spatial Pattern of Sporting Achievement Level in China and its Relation with Economic Development

Ping Zhang [1*], Lupeng, Guan [1], Peter M. Atkinson [2]

[1] Geo-Exploration Science & Technology College, JiLin University, Changchun, China
Email: p.zhang1020@gmail.com

[2] Geography and Environment, University of Southampton, Highfield, Southampton, UK
Email: P.M.Atkinson@soton.ac.uk

* Corresponding author: Ping Zhang

## 1. Introduction

The spatial distribution of the competitive sports achievement level in China is inhomogeneous, with vast differences between different regions. For example, there is a large difference between the south and the north of China, delimited by the Qinling Huaihe River. The north has an advantage in physical and mixed sports; while the south has an advantage in technical sports. For the Northeast, it is culturally a great advantage to have sporting talent, but this is not so for the Southwest and Northwest. Therefore, it is interesting to analyze the spatial pattern of competitive sports achievement level and the impact of the economy on sporting achievement in China.

Here, competitive sports refer to sporting activities in which the goal is to win in matches, either in teams or competing as individuals. Competitive sports have attracted research (e.g. Lu and Bai, 2005; Wang and He, 2009; Tao and Lin, 2011; Wu and Zhao, 2012; Zhang and Yu, 2013). Among these, three (Lu and Bai, 2005; Wang and He, 2009; Tao and Lin, 2011) were about competitive sports achievement and two were about sporting talent and Olympic medals. So the use of spatial statistical methods and mathematical models in research on competitive sports achievement level in China is rare.

## 2. Methods

### 2.1. Getis's $G_i$ statistic

Getis's $G_i$ statistic, developed by Getis and Ord (1992), is a multiplicative measurement of the overall spatial association of values which fall within a critical distance of each other. It can be used as a method of detecting hotspots, and can be expressed as follows:

$$G_i = \frac{\sum_{j}^{n} w_{ij}(d)\, y_i - w_i \bar{y}}{S\left\{[nS_{1i} - w_i^2]/(n-1)\right\}^{1/2}} \qquad (1)$$

where $S$ is the standard variance of the sports level; when the distance from place $j$ to $i$ is within distance $d$, then $w_{ij}(d)=1$; otherwise $w_{ij}(d)=0$. The higher the value of $G_i$, the greater the influence of place $i$ at a given distance $d$, indicating that place $i$ is a hotspot of the region.

## 2.2. Environmental variables

Here, three kinds of environmental variables have been identified and used based on the previous work (Wang and He, 2009), including economic environmental variables, anthropogenic environmental variables and physical environmental variables (Table 1). Economic environmental variables include (i) per capita GDP; (ii) per capita consumption level; (iii) sports lottery funding; (iv) number of sports projects. Anthropogenic environmental variables include (i) population density; (ii) number of professional athletes; (iii) number of full-time sports instructors; (iv) number of people employed in sports schools, colleges and universities; (v) number of public guidance personnel; (vi) sports ground area; (vii) number of sports units. Physical environmental variables include (i) elevation; (ii) slope; (iii) annual mean temperature; (iv) annual mean precipitation.

## 2.3. Spatial regression model

Spatial regression models are regression models which include a term to deal with spatial dependence in the residuals, which may arise from several sources, including unobservable latent variables that are spatially correlated (Altman et al 2004). Here, a generalized linear mixed model (GLMM) incorporating a variogram model was used as a spatial regression model. Specifically, the Poisson log-linear mixed model was used. Suppose that, given the random effects $\alpha$, the counts $y_1 \ldots y_n$ are conditionally independent such that:

$$y_i \mid \alpha \sim \text{Poisson}(\lambda_i) \tag{2}$$

$$\log(\lambda_i) = x_i'\beta + z_i'\alpha \tag{3}$$

where $x_i'$ and $z_i'$ are known vectors, $\beta$ is a vector of unknown parameters (the fixed effects), $\lambda_i$ is the expected sports level during the given interval.

The Poisson log-linear model with a random intercept fitted through the PQL estimation method used here can be written:

$$\ln(\lambda_i) = \beta_0 + x_i\beta_i + b_i \tag{4}$$

Table 1. Original scale and final scale of environmental variables.

| Category | Description of variables | Resolution and manipulation |
|---|---|---|
| Economic environmental variables | per capita GDP | County level; Interpolated into 5km×5km |
| | per capita consumption level | Province level; Allocated to county level; Interpolated into 5km×5km |
| | sports lottery funding | Province level; Allocated to county level; Interpolated into 5km×5km |
| | number of sports projects | Province level; Allocated to county level; Interpolated into 5km×5km |
| Anthropogenic environmental variables | population density | 1042m×1042m; Resample into 5km×5km |
| | number of professional athletes | Province level; Allocated to county level; Interpolated into 5km×5km |
| | number of full-time sports instructors | Province level; Allocated to county level; Interpolated into 5km×5km |
| | number of people employed in sports schools etc | Province level; Allocated to county level; Interpolated into 5km×5km |
| | number of public guidance personnel | Province level; Allocated to county level; Interpolated into 5km×5km |
| | number of sports units | Province level; Allocated to county level; Interpolated into 5km×5km |
| | area of school sports ground | Province level; Allocated to county level; Interpolated into 5km×5km |
| Physical environmental variables | elevation | 1042m×1042m; Resample into 5km×5km |
| | slope | 1042m×1042m; Resample into 5km×5km |
| | annual mean temperature | 500m×500m; Resample into 5km×5km |
| | annual mean precipitation | 500m×500m; Resample into 5km×5km |

where $\lambda_i$ is the sports level, $\beta_0$ and $\beta_i$ are the unknown parameters for the fixed effects, $x_i$ are the environmental variables, $b_i$ are the random effects with distribution assumption:

$$b_i \sim N(0, \sigma^2) \tag{5}$$

Equation (5) means that the random effects of $b_i$ are normally distributed with mean 0 and variance $\sigma^2$.

GLMM was chosen here because it allowed for spatial correlation structure in the residuals through its spatial random effects term (Fuller et al., 2010). The spatial random effects term $b_i$ is similar to the residuals (error term) in classical linear models (equation 4). Here, a geostatistical (variogram) model was used to represent the spatial autocorrelation term in the residuals $r_i$, written as:

$$r_i \sim N(\mu, \sigma^2) \tag{6}$$

$$\sigma^2 = I\sigma_1^2 + F\sigma_2^2 \tag{7}$$

$$F = \exp(-d_{ij} / \rho) \tag{8}$$

where, the residuals $r_i$ of the GLMM (equation (4) are distributed normally with mean $\mu$ and variance $\sigma^2$), $\sigma_1^2$ is the nugget of the residuals' semi-variance, $\sigma_2^2$ is the sill, $\rho$ is the range, $d_{ij}$ is the lag distance, $I$ is the adjusted coefficient.

## 3. Results

### 3.1. Spatial pattern analysis of competitive sports achievement level

We consider first the general Getis G statistic. For the general Getis G statistic, the $z$ value was 2.38, with a p value of 0.017. This means that sports achievement level at the province level has a spatially clustered (non-random) distribution. The observed value of G was 0.04. Since the general G statistic is greater than zero, this indicates that hot spots for sports level exist in China.

We now consider the local Getis G statistic For the local Getis G statistic, $z$ values greater than 1.96 were found in Yangtze Delta Region, containing Jiangsu, Zhejiang, Anhui; $z$ values greater than 1.65 and less than 1.96 were found in the areas along Bo Hai; $z$ values less than -2.58 were found in western parts including Xinjiang, Tibet, Qinghai and Chuanyu (Figure 1). Thus, sports level in the Yangtze Delta Region is the highest in China, forming a hot spot for sports level.

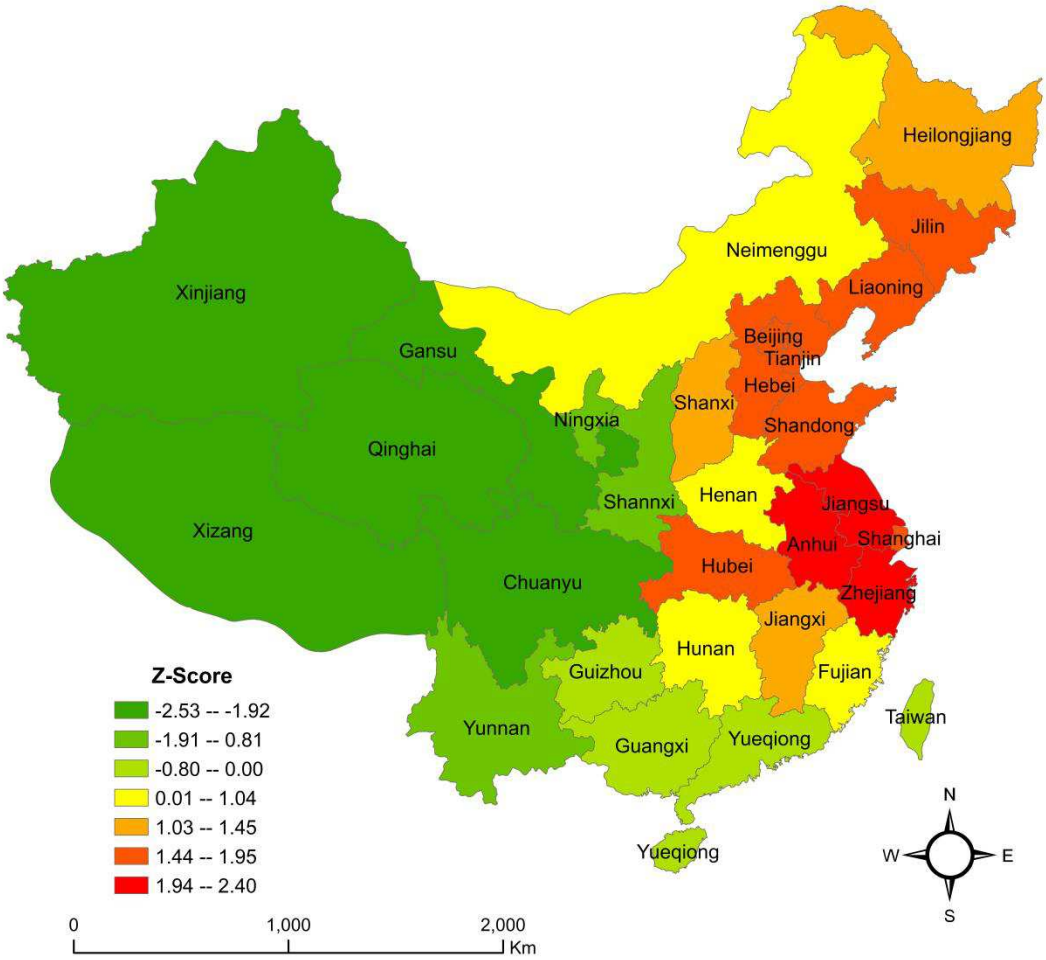



Figure1. Spatial distribution of local Getis G statistic for competitive sports level in China.

## 3.2. Environmental variables effects analysis

Four environmental variables have effects on the competitive sports achievement level at province level in China (Table 2). Specifically, these are per capita GDP, sports ground area, number of sports projects and number of people employed in sports schools, colleges and universities because their P-values are less than or equal to 0.05. In particular, the estimated coefficients of these four environmental variables are all not less than zero, which means all of them are positively correlated to the competitive sports achievement level. Sports industries can improve with the increase in per capita GDP, sports ground area, number of sports projects, number of people employed in sports schools, colleges and universities in a country. Certainly, other environmental variables could influence the development of competitive sports achievement level, but their significances are not as notable as these four variables especially in the sense of statistics.

Table 2. Effects of environmental variables on competitive sports level in China 2008.

| Variable ( $X_i$ ) | Estimated Coefficient ( $\beta_i$ ) | Std. Error | DF | T-value | P-value |
|---|---|---|---|---|---|
| (Intercept) | -0.460 | 0.353 | 23 | -1.303 | 0.192 |
| per capita GDP | 0.000 | 0.000 | 23 | 2.621 | 0.012 |
| sports ground area | 0.000 | 0.000 | 23 | 2.404 | 0.016 |
| sports projects | 0.001 | 0.000 | 23 | -2.650 | 0.024 |
| people employed in sports | 0.033 | 0.015 | 23 | 2.152 | 0.042 |

## 4. Conclusion

This paper uses the integration of the percentage of world champion, football teams' ability, basketball teams' ability, sports ability in National Games, citizen physique level, sports subjects' evaluation points and the percentage of sports ground to measure the competitive sports achievement level in China. Through applying the Getis G statistic, it was shown that high values of competitive sports achievement are spatially clustered in China. Moreover, the local Getis G statistic showed that a hotspot for competitive sports achievement level exists in China. In addition, through applying a Poisson GLMM model at the pixel resolution of 5000 m by 5000 m, associations between competitive sports achievement level and three classes of environmental variables were found, with four environmental variables having a significant effect, including per capita GDP, sports ground area, number of sports projects and number of people employed in sports schools, colleges and universities. These findings could shed light on the spatial pattern of competitive sports achievement level and its influencing environmental variables at the province scale. The methods used here provide a prototype for exploring the spatial pattern of competitive sports achievement level and its associations with environmental variables for other countries.

## References

Fuller TL, Saatchi SS, Curd EE, Toffelmier E, Thomassen HA, Buermann W, DeSante DF, Nott MP, Saracco JF, Ralph CJ, Alexander JD, Pollinger JP, Smith TB, 2010, Mapping the risk of avian influenza in wild birds in the US. BMC Infectious Diseases, 10: 187.

Getis A and Ord JK, 1992, The analysis of spatial association by use of distance statistics. Geographical Analysis, 24: 189-206.

Lu F and Bai J, 2005, Athletic sports' spatial-temporal regional segregation and factors analysis in China. Journal of Xuchang University, 24(5): 69-72. (In Chinese)

Tao S and Lin L, 2011, Resource allocation in social transition: study on regional differentiation of competitive sports. China Sport Science, 31(7): 3-7. (In Chinese)

Wang L and He Q, 2009, Research on spatial distribution and influencing factors of provincial athletic sports level in China. Journal of Beijing Sport University, 32(10): 109-112. (in Chinese)

Wu M and Zhao Y, 2012, Analysis on Chinese Olympic medals spatial distribution based on Moran's I index. China Sport Science and Technology, 48(5): 3-9. (In Chinese)

Publications from the **GEOINFO-SERIES**:

1. Experiential Realism and its Applications to Geographic Space
   *Compiled by Irene Campari and Andrew Frank*
2. Temporal Data in Geographic Information Systems (2nd print)
   *Compiled by Andrew Frank, Werner Kuhn and Peter Haunold*
3. Geo-Information Management Systems- A Major Public Issue and its Educational Challenges
   *Andrew U. Frank and Irene Orchard*
4. Geographic Information Systems - Materials for a Post-Graduate Course
   Vol. 1: Spatial Information
   *Andrew U. Frank, Editor*
5. Geographic Information Systems - Materials for a Post-Graduate Course
   Vol. 2: GIS Technology
   *Andrew U. Frank, Editor*
6. Geographic Information Systems - Materials for a Post-Graduate Course
   Vol. 3: GIS Organization
   *Andrew U. Frank, Editor*
7. Semantics of Geographic Information
   *Werner Kuhn*
8. Spatialization: Spatial Metaphors for User Interfaces (2nd print)
   *Werner Kuhn and Brad Blumenthal*
9. COSIT'95 Doctoral Consortium
   *Compiled by Werner Kuhn and Sabine Timpf*
10. Hierarchical Spatial Reasoning: Theoretical Consideration and its Application to Modeling Wayfinding
    *Adrijana Car*
11. Aufdeckung von numerischen Problemen in geodätischer Software
    *Christine Goldenhuber*
12. Gofer as used at GeoInfo/TU Vienna
    *Andrew U. Frank, Werner Kuhn, Werner Hölbling, Hartmuth Schachinger, Peter Haunold*

13. Hierarchical Structures in Map Series
    *Sabine Timpf*
14. Organisation des Katasters – Ziele, Grundsätze und Praxis
    *Christoph Twaroch*
15. Die Modellierung eines Grundbuchsystems im Situationskalkül
    *Steffen Bittner*
16. Multi-Agency Databases to Manage Geographic Information
    *Andrew U. Frank, Martin Raubal, Maurits van der Vlugt (Editors)*
17. Technical Concept for Pay-per-Use in Geomarketing Services
    *Peter Gustav Wenzl*
18. Transformation und Inspektion mentaler Umraumrepräsentationen: Modell und Empirie
    *Annette von Wolff*
19. Geographical Domain and Geographical Information Systems
    *Stephan Winter (Editor)*
20. Unified Behavior of Spatial Data Representations
    *Stephan Winter*
21. PANEL-GI Compendium, a Guide to GI and GIS
    *Andrew Frank, Martin Raubal, Maurits van der Vlugt (Editors)*
22. A Cost Oriented Approach to Geodetic Network Optimisation
    *Martin Staudinger*
23. Rough Location
    *Thomas Bittner*
24. An agent-based motel of reality in a cadastre
    *Steffen Bittner*
25. Formalisierung von Gesetzen
    *Gerhard Navratil*
26. An Agent-Based Model for Quantifying the Economic Value of Geographic Information
    *Alenka Krek*
27. Wayfinding in Build Environments
    *Martin Raubal*

28a. Proceedings of the ISSDQ ´04 Volume 1
    *Andrew U. Frank, Eva Grum*

28b. Proceedings of the ISSDQ ´04 Volume 2
*Andrew U. Frank, Eva Grum*

29. Festschrift zum Simon von Stampfer Symposium
*Johanna Brückl, Gerhard Navratil (Editors)*

30. Influences of Technology, Law, and Usability on Data Quality
*Gerhard Navratil*

31. Proceedings of the IWWPST ´05
*Andrew U. Frank (Editor)*

32. Combinatorial Optimization in Geography
*Takeshi Shirabe*

33. Proceedings of the IWWPST ´06
*Andrew U. Frank (Editor)*

34. Bewegter Planungprozess
*Christine Rottenbacher*

35. Route-Choice Strategies for Shared-Ride Trip Planning in Geosensor-Networks
*Christian Gaisbauer*

36. Wayfinding in GIS: Formalization of Basic Needs of a Passenger When Using Public Transportation
*Elissavet Pontikakis*

37. Lifestyles - A Paradigm for the Description of Spatialtemporal Databases
*Damir Medak*

38. Sandbox Geography
*Florian Twaroch*

39. Proceedings of the Colloquium for Andrew U. Frank's 60th Birthday
*Gerhard Navratil (Editor)*

40. Extended Abstract Proceedings of GIScience 2014
*Matt Duckham, Paolo Fogliaroni, Gerhard Navratil, Edzer Pebesma, Kathleen Stewart (Editors)*

The books are available from:

Department of Geodesy and Geoinformation
Vienna University of Technology
Gusshausstraße 27-29/120.2
A-1040 Vienna, Austria
fax: ++43-1-58801-12799
Email: gruber@geoinfo.tuwien.ac.at
http://www.geoinfo.tuwien.ac.at/

Except:

Nr. 27: available at the Universität Münster, Institut für Geoinformatik
Nr. 29: available at the Institute of Geodesy and Geophysics, Vienna University of Technology